Universidade do Minho
Escola de Ciências

Dora Sofia Martins Henriques

Employment of whole genome resequencing to reveal the evolutionary history and to develop molecular tools for Western European honey bees (*Apis mellifera* subspecies)

janeiro de 2018

Employment of whole genome resequencing to reveal the evolutionary history and to develop molecular tools for Western European honey bees (*Apis mellifera* subspecies)
Dora Sofia Martins Henriques
UMinho | 2018

FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA
POPH QUALIFICAR É CRESCER.
QREN QUADRO DE REFERÊNCIA ESTRATÉGICO NACIONAL PORTUGAL 2007.2013
Governo da República Portuguesa
UNIÃO EUROPEIA Fundo Social Europeu

Universidade do Minho
Escola de Ciências

Dora Sofia Martins Henriques

Employment of whole genome resequencing to reveal the evolutionary history and to develop molecular tools for Western European honey bees (*Apis mellifera* subspecies)

Tese de Doutoramento em Biologia Molecular e Ambiental
Especialidade em Evolução, Biodiversidade e Ecologia

Trabalho efetuado sob a orientação da
Professora Doutor Maria Alice da Silva Pinto
do
Professor Doutor Filipe José Oliveira Costa
e do
Professor Doutor Matthew Thomas Webster

janeiro de 2018

DECLARAÇÃO

NOME: Dora Sofia Martins Henriques

ENDEREÇO ELECTRóNICO: dorasmh@gmail.com

TÍTULO DA TESE:

Employment of whole genome resequencing to reveal the evolutionary history and to develop molecular tools for Western European honey bees (*Apis mellifera* subspecies)

ORIENTADORA:

Professora Doutor Maria Alice da Silva Pinto

CO-ORIENTADORES:

Professor Doutor Filipe José Oliveira Costa

Professor Doutor Matthew Thomas Webster

ANO DE CONCLUSÃO: 2018

DOUTORAMENTO EM: Biologia Molecular e Ambiental
ESPECIALIDADE EM: Evolução, Biodiversidade e Ecologia

Universidade do Minho, 31 / 01 / 2018

Assinatura: Dora Sofia Martins Henriques

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração da presente tese. Confirmo que em todo o trabalho conducente à sua elaboração não recorri à prática de plágio ou a qualquer forma de falsificação de resultados.

Mais declaro que tomei conhecimento integral do Código de Conduta Ética da Universidade do Minho.

Universidade do Minho, 31 de Janeiro de 2018

Nome completo: Dora Sofia Martins Henriques

Assinatura: Dora Sofia Martins Henriques

Dissertação apresentada à Escola de Ciências da Universidade do Minho para obtenção do grau de Doutor em Biologia Molecular e Ambiental.

Este trabalho foi realizado no Centro de Investigação de Montanha do Instituto Politécnico de Bragança, sob a orientação da Professora Doutora Maria Alice da Silva Pinto, no Departamento de Bioquímica Médica e Microbiologia da Uppsala Universitet, sob co-orientação do Professor Doutor Matthew Thomas Webster e no Centro de Biologia Molecular e Ambiental da Universidade do Minho, sob co-orientação Professor Doutor Filipe José Oliveira Costa

A sua execução foi financiada pela Bolsa de Doutoramento SFRH/BD/84195/2012, atribuída pela Fundação para a Ciência e a Tecnologia (FCT). Contou ainda com os recursos disponibilizados pelo projecto BeeHope: 2013-2014 BiodivERsA/FACCE-JPI (joint call for research proposals, with the national funders FCT, Portugal, CNRS, France, and MEC, Spain) e do projeto COMPETE/QREN/EU (PTDC/BIA-BEC/099640/2008).

DEDICATORIA

“Cada um que passa na nossa vida, passa sozinho, pois cada pessoa é única e nenhuma substitui outra. Cada um que passa na nossa vida, passa sozinho, mas não vai só nem nos deixa sós. Leva um pouco de nós mesmos, deixa um pouco de si mesmo”.

“Every one that goes in our life passes alone, because each person is unique and no other substitutes. Every one that goes in our life passes alone, but will just not leave us alone. It takes a bit of ourselves, leaves a bit of himself”.

Antoine De Saint Exupery

Aos meus pais ,João e Leonor,

irmãos, Marisa e Marco,

e ao meu jovem sobrinho André

ACKNOWLEDGMENTS

Doing my PhD is an old dream of mine, that sends me to a time where I didn't know what a doctorate or PhD was (if someone could know what it is before actually trying do it). These have been the most intense 4 years and 4 months of all my life, that I could recall that is. It has been a peculiar journey, with ambiguous feelings along the way, most of the time because I thought I should do more or something different, or perhaps because I had a sense of uncertainty regarding the future. My generation has so many benefits, but I always thought that the subject of employment would be easier to dealt with. I was rarely driven by economic factors, I always thought we should do what we like because we would better that way. Looking back now, I am certain that nothing could have been different (perhaps those sentiments of uncertainty may have been unnecessary) or, at least, the path could not have been any other way. It was a long and happy journey, although it is also a joy to be able to complete it, because nothing makes sense to initiate if it is not to complete it. And everything is just a challenge so that other challenges can wait for us, which I also hope to happen. It was a path filled with challenges (who knew I would learn to program and even more...enjoy doing it also?). Or go to Sweden, right from the start of the PhD. Who would say that I would come to Bragança? Yes, it was a literal journey in a lot of lands, a lot of countries (Isn't it marvelous travelling? Living in another country? Knowing other cultures? I honestly hope the world keeps coming closer and closer) and, more importantly, a lot of people. To all of you, to whom I must thank. The words will not be as pretty as I would like (that thing called lack of time!!!) nor will even show everything I feel.

In the first place, because without her the PhD would not be possible, I would like to thank my advisor, Professor Alice Pinto. I would like to thank her mainly because she never ceased to believe in me and because she always had a word to calm down so many of my doubts. Working with you has been an adventure, in a good way. I have been learning and I know that I still have a lot to learn. I must say I have a feeling of admiration and friendship for you, for all the reasons that I find hard to express only through writing.

To Professor Filipe Costa, for the support in this thesis and particularly for the support with all the bureaucracies that involve a thesis.

To Matthew Webster for giving me the chance to go to Sweden, it was the contact with a reality different from the one I knew. I learned a lot; to analyze genomes, to do software

programming, the English language, a new culture. Thank you for all the support, both professional and personal.

To Professor José Rufino, for all the computer support and all of his patience.

To Miguel Carneiro, for the interesting discussions we had about the different themes of the thesis.

To my parents (João and Leonor) and my brother and sister (Marco and Marisa) and to the new members of my family (Jacinta and André). I dedicate this thesis to all of you, because without your support in the realization of all of my dreams, even the ones that may appear to be less rational, I would not be here. You are my safe haven, my base, my strength, my calmness. I love you with all my heart.

To my grandmother Maria, uncles and aunts (Tila, Miguel, Catarina, Carlos, Domingos, Lurdes, José, Maria, Albertino and António), my cousins (Daniela, David, João, Micael, Filipe, Inês, Maria Amélia and Lurdes) you know that you are special and know how hard it is for me to not participate more often in our family meetings. We share genes and we share a life. You are a part of my life, you are a part of me.

To my goddaughter, Carolina Crespo, for having inspiring moments together. You are beautiful and I hope I could be even more present in your life.

To Artur Gonçalves, for all the support, all the calmness and patience and, especially, for all the magic you put into my life. A lot of adventures still wait for us, because, lately, what has been waiting for you is an unfinished thesis demanding from you a great dose of patience. Thank you for all you have been doing. Without you wouldn't be possible.

To Carlos and Emília Gonçalves for the kindness, support and precious help, mainly during the last month of this long journey.

To my friends of the bee club, which are like brothers and sisters to me. They all have been an adventure that, now that I think of it, is starting to have a lot of names. To Cátia Neves (Kate Snow), my "bifásica" ("two-phase girl"), with whom I fell in love and have shared so many years, so many adventures. Thank you for always being by my side, for always knowing the best and worst of me, for caring and for being who you are, my "sister-arm". To Júlio Chávez-Galarza for our special friendship that was formed while we worked together. You will always be like a brother with whom I spend Christmas. To my eternal freshman girl, Helena Ferreira, you have grown so much all these years, likewise our friendship, and it is good to being able to witness that and I hope

it continues that way; you know exactly who showed me the poem! To my beloved Andreia Quaresma, my "*Dipsas Albifrons",* if time is relative so is friendship; have you noticed that not even a year has passed since we know each other? And, sometimes, it feels like we have known each other our whole lives. I hope life still brings us a lot of adventures, because I like you. To Ana Rita Lopes, the "last acquisition" of the club, I must say that it seems to me that club became perfect now. Thank you for all your help and good mood. And, of course, to our dear beekeeper Paulo Ventura, for always being there, for being good-natured and for your bee knowledge. Now the bee club is not only for those who work with bees, but also for those that don't even know they belong to that club. To Miguel Vaz-Pinto, for your friendship, for all the moments that we spent together, I miss you so much. To Fernando Pérez for being a special friend (yes, friend!), for being a good listener, for knowing when I need to go for a walk or a coffee at 6 am. The club without you is not the same. To Ângelo Sil, for your good mood and good conversations. To my dear Eric Carvalho, Marisa Barroso e Dioguinho, for making our lives more joyful and for being such loyal friends. I miss you and miss the times we spent together. To Irene Muñoz for your help and the time that we spend together not only working, but also having fun. See you around!! To Melanie Parejo, for your friendship, for your help and for your companionship. You were one of the people that I most enjoyed knowing and we already shared so much together, not just papers and I hope it continues to be that way. To Keith Browne, it was good to having you here and showing you a little of our culture. Thank you for also helping me with my English and for being a good friend. Come again any time you want or we will visit you there.

To Fernando Cardoso da Cunha, for everything we've been through and for all the support that you gave me and my family. I wish you the best that life has to offer you. To my grandmother, Ana Cardoso, that will always be in my heart and to Ana Lurdes for your friendship and support that were essential. To your children too, of course, Maria, Ana and Carolina, who I like so much and I know that you will have a bright future waiting for you. To Iolanda Quintela for the friendship, words of encouragement and help. To Fausto Montenegro, for the friendship, for all the dinners always accompanied by a good Douro wine. To Filipe Ribeiro Torres for the cultural dinners, always a time well spent, and for sharing so much of your knowledge. To Ângelo Quintela for your sympathy. To Diana Bastos and Luís Martins for being exemplary friends and for always being present.

To my longtime friends, Melissa Rodrigues, my beloved roommate, Nataxa Freitas, Inês Rondão e Andreia Raimundo with whom I discovered science, to Daniela Ascensão, Ana Rita Nunes, Alice Mendes, Carla Pinto, Ana Silva, Lídia Birolo, Sónia Zacarias, Jéssica Alvarado, Ana Afonso and Rui Santos for your friendship that goes beyond time, especially because time is something that I have not been dedicating to you, but I hope to make up for it anytime soon.

To my housemates, Telma Teixeira and Diana Rabasquinho, for the crazy adventures and the amazing friendship. Every time it is snowing, I think about you. To Sónia Macedo, for your love and affection and the strength that you express. From now on, I can accompany you a little more. To the Marie Becker and Louise Christiansen for make me feel like I was home, I miss you so much and I would like to spend more time with both of you. I can assure you, that I will take you and our adventures with me for the rest of my life.

To Andreas Wallberg for everything that you taught me, that highly contribute for my professional development and for my thesis. Thank you for all the great moments and for showing me how amazing is Sweden. And for course for the abstract!!

To my team colleagues, Ronald Nelson, Martin Schmid and Anna Olsson for the interesting lunches, for the coffees and companionship. Ann your presence is like a sunshine in a dark day.

To Fabiana to whom I am really greatfull for helping me when I needed. What you did for me will stay forever in my heart and I hope to see you and Piu soon.

To me colleagues in the department for making my journey so special: to Jonas Berglund, thank you for the bike and the support in the resolution of my problems; to Freyja Imsland for being so different and for all the help; to Matteo Bianchi for being a good colleague at the office; to Sam Barsh and Daniela Hahn for all the adventures; to Angela Fuentes-Pardo, Vicky, Johanna, Sara Negro, Ann Staiger, Jagoda Jablonska, Marcin Kierczak, Sangeet Lamichhaney and Sergei Abramov for making my journey in Sweden so much fun.

To, my colleagues from Sweden, André Tiso Lobato and Juan Pulgarin for all the good moments we spent together.

To Sara Martins: now Braga will never be the same without you, thank you for being an excellent colleague in programming classes and always make your home available for me, but mainly for your friendship.

To Almir Smith and the yôga team for teaching me and for being part of my relax moments when I most needed.

The doctorate time has been rich and interesting...

"Conversely, rich and interesting content is capable of shortening and quickening the hour and even the actual day; on a large scale, though, it endows the course of time with breadth, weight and solidity, so that eventful years pass much more slowly than those poor, empty light years which the wind blows before it, and which fly away."

Thomas Mann "The Magic Mountain"

AGRADECIMENTOS

Fazer o doutoramento consiste num sonho antigo que remete para o tempo onde eu ainda não sabia o que era o doutoramento (se é que se sabe o que é antes de se tentar fazer). Estes têm sido os 4 anos e 4 meses mais intensos de sempre, pelo menos de que tenho memória. Tem sido um percurso peculiar, com sentimentos ambíguos ao longo do caminho a maioria das vezes por achar que devia fazer mais ou diferente, ou talvez pelo sentimento de incerteza em relação ao futuro. A minha geração tem tantos benefícios, mas eu achava que isto do emprego seria mais fácil, raramente me movi por razões económicas, sempre achei que devíamos fazer o que gostamos, porque assim seriamos melhores. Olhando para trás tenho a certeza que nada poderia ter sido diferente (talvez esses sentimentos de insegurança sejam desnecessários) ou pelo menos o caminho não podia ter sido de outra maneira. Foi um percurso longo e feliz, claro que é feliz também poder terminá-lo, porque nada faz sentido iniciar se não for para terminar. E tudo não passa de um desafio para que depois outros maiores nos esperem, assim eu espero também. Foi um percurso de desafios, quem diria que iria aprender a programar e mais...a gostar de programar. Ou ir para a Suécia, bem começando da base. Quem diria que viria para bragança. Sim, foi um percurso com muitas terras, muitos países (Não é maravilhoso viajar? Viver noutro pais? Conhecer outras culturas? Sinceramente espero que o mundo se continue a aproximar) e principalmente muita gente. A todos vocês a quem tenho de agradecer. As palavras não serão tão bonitas como eu gostaria (essa coisa da falta de tempo!!!) nem sequer irão mostrar tudo aquilo que sinto.

Em primeiro lugar, porque sem ela o doutoramento não seria possível, gostaria de agradecer à minha orientadora a professora Alice Pinto. Queria agradecer principalmente por nunca ter deixado de acreditar em mim e por ter tido sempre uma palavra que acalmou tantas dúvidas. Trabalhar consigo tem sido uma aventura, no bom sentido, tenho aprendido e sei que ainda tenho muito por aprender. Devo dizer que nutro um sentimento de admiração e de amizade por si por todas as razões que me é difícil escrever.

Ao professor Filipe Costa pelo apoio nesta tese e em particular com todas a burocracias que envolvem fazer uma tese.

Ao Matthew Webster por me ter dado a oportunidade de ir para Suécia, foi o contacto com uma realidade diferente daquela que eu conhecia. Aprendi muito; a analisar genomas, a programar, o inglês,uma nova cultura. Obrigada por todo o apoio não só profissional, como também pessoal.

Ao professor José Rufino por todo o suporte informático e por toda a paciência.

Ao Miguel Carneiro pelas discussões interessantes sobre os diferentes temas da tese.

Aos meus pais (João e Leonor) aos meus irmãos (Marisa e Marco) e anos novos membros da família (Jacinta e André). A vós dedico esta tese, porque sem o vosso apoio na concretização de todos os meus sonhos, mesmo os que possam parecer menos racionais, eu não estaria aqui. Vocês são o meu porto seguro, vocês são a minha base, a minha força, a minha calma. Amo-vos com todo o meu coração.

Aos meus; à minha avó Maria, tios (Tila, Miguel, Catarina, Carlos, Domingos, Lurdes, José, Maria, Albertino, António), primos (Daniela, David, João, Micael, Filipe, Inês, Maria Amélia, Lurdes) vocês sabem que são especiais e sabem o quanto me custa muitas vezes não participar nas reuniões familiares. Partilhamos genes e partilhamos uma vida. Fazem parte da minha vida, fazem parte de mim.

À minha afilhada Carolina Crespo por termos juntas momentos inspiradores. És linda e espero poder estar ainda mais presente na tua vida.

Ao Artur Gonçalves, por todo o apoio, por toda a calma e paciência e principalmente por toda a magia que colocas na minha vida. Ainda muitas aventuras nos esperam, porque ultimamente o que te tem esperado é uma tese inacabada exigindo de ti uma boa dose de paciência. Obrigada por tudo o que tens feito. Sem ti isto não seria possível.

Carlos e Emília Gonçalves pela simpatia e apoio e por terem sido uma ajuda preciosa, principalmente neste último mês.

Aos meus irmãos e amigos do clube da abelha. Tem sido uma aventura que pensando bem começa a ter muitos nomes. À Cátia Neves a minha bifásica por quem me apaixonei e que temos partilhado tantos anos, tantas aventuras. Obrigada por estares sempre a meu lado, por saberes o melhor e pior de mim, por te preocupares e simplesmente por seres quem tu és a minha braço-irmã. O Júlio Chávez-Galarza pela nossa amizade especial que formamos enquanto trabalhávamos juntos, serás sempre como irmão com quem eu passo o Natal. À minha eterna

caloira, Helena Ferreira, tens crescido tanto estes anos, assim como a nossa amizade, e é bom poder assistir e espero que assim continue, sabes bem quem me mostrou o poema! À minha querida Andreia Quaresma a minha "*Dipsas albifrons*" se o tempo é relativo a amizade também, já viste que ainda nem há um ano nos conhecemos? E às vezes parece que nos conhecemos a vida toda. Espero que a vida nos traga ainda muitas aventuras, porque gosto de ti. À Ana Rita Lopes a "última aquisição" do clube, devo dizer que me parece que assim o grupo ficou perfeito, obrigada por toda a tua ajuda e boa disposição. E claro ao nosso querido amigo apicultor Paulo Ventura, por estares sempre lá, por seres bem-disposto e pelo teu conhecimento sobre as abelhas. Agora o clube da abelha não é só para quem trabalha com abelhas, mas também aqueles que talvez nem saibam que pertençam a esse clube. Ao Miguel Vaz-Pinto pela tua amizade, por todos os momentos que passamos juntos, tenho muitas saudades tuas, ao Fernando Pérez por seres um amigo (sim, amigo!) especial, por saberes ouvir, por saberes quando preciso de uma caminhada ou de um café às 6h da manhã, o clube sem ti não é a mesma coisa. Ao Ângelo Sil pela boa disposição e sempre uma boa conversa. Aos meus queridos Eric Carvalho, Marisa Barroso e Dioguinho, por terem feito a nossa vida bem mais alegre e por serem uns amigos tão leais, tenho saudades vossas e do tempo que passámos juntos. À Irene Munõz pela ajuda e pelo tempo que passámos juntas não só a trabalhar, mas também de diversão, iremos vernos por aí!! À Melanie Parejo pela tua amizade, pela tua ajuda pelo teu companheirismo. Foste uma das pessoas que mais gostei conhecer e já partilhámos muitas coisas juntas, para além de papers e espero que continue a ser assim. Ao Keith Browne, foi muito bom ter-te aqui e mostrar-te um pouco da nossa cultura. Obrigada também por me ajudares no Inglês e por seres um bom amigo, volta sempre ou nós iremos aíi.

Ao Fernando Cardoso Cunha, por tudo o que passámos juntos e por todo o apoio que me deste a mim e à minha família, desejo-te o melhor que a vida tem para te dar. A avó Ana Cardoso que estará sempre no meu coração, à Ana Lurdes pela tua amizade e apoio que foram essências, claro que também as tuas filhotas Maria, Ana e Carolina, de quem eu gosto tanto e sei que vais ter um futuro brilhante à tua espera. À Iolanda Quintela pela amizade, palavras de incentivo e ajuda. Ao Fausto Montenegro, pela amizade, por todos os jantares sempre acompanhados por um bom vinho do Douro. Ao Filipe Ribeiro Torres pelos jantares culturais sempre bem passados e por partilhar tanto conhecimento. Ao Ângelo Quintela pela sua simpatia. À Diana Bastos e ao Luís Martins por serem amigos exemplares e estarem sempre presentes.

Aos meus amigos de sempre Melissa Rodrigues, a minha querida colega de quarto, Natacha Freitas , Inês Rondão e Andreia Raimundo com quem descobri a ciência, Daniela Ascensão, Ana Rita Nunes, Alice Mendes, Carla Pinto, Ana Silva, Lídia Birolo, Sónia Zacarias, Jéssica Alvarado, Rui Santos e minha querida Ana Afonso pela vossa amizade que vai para além do tempo, até porque tempo é algo que não vos tenho dedicado, mas espero poder compensar em breve.

Às minhas colegas de casa, à Telma Teixeira, Diana Rabasquinho, foram tantas as aventuras e tão grande a amizade agora cada vez que neva penso em vocês, à Sónia Macedo, pelo amor e carinho e a força que transmites, a partir de hoje posso acompanhar-te um pouco mais, à Marie Becker e Louise Christiansen por me terem feito sentir em casa, tenho muitas saudades e gostava de estar com vocês, as nossas aventuras vão sempre permanecer no meu coração.

Ao Andreas Wallberg por tudo o que me ensinou que foi essencial para o desenvolvimento da minha tese. Obrigada também pelos bons momentos e por me mostrares a Suécia. E claro pelo abstract em Sueco!

Aos meus colegas de grupo Ronald Nelson, Martin Schmid e Anna Olsson pelos almoços interessantes, pelos cafés e pela companhia. Anna a tua presença ilumina um dia escuro.

À Fabiana a quem devo muito por me ter ajudado quando eu mais precisei. O que fizeste por mim ficará para sempre no meu coração, espero que brevemente te possa ver a ti e ao piu.

Aos meus outros colegas de departamento por terem feito a minha estadia tão especial, ao Jonas Berglund, obrigada pela bicicleta e pelo apoio da resolução de problemas, à Freyja Imsland por me ajudares e seres especial. Ao Matteo Bianchi por ser um bom colega de gabinete, ao Sam Barsh e Daniela Hahn por todas as aventuras, à Angela Fuentes-Pardo, Vicky, Johanna, Sara Negro, Ann Staiger, Jagoda Jablonska, Marcin Kierczak, Sangeet Lamichhaney e Sergei Abramov por fazerem a vida na suécia bem mais divertida.

Ao meus outros colegas da Suécia, André Tiso Lobato e Juan Pulgarin pelos bons momentos passados juntos.

À Sara Martins, agora Braga não é o mesmo sem ti, obrigada por teres sido uma excelente colega nas aulas de programação e pela tua casa, mas principalmente, obrigada pela tua amizade.

Ao Almir Smith e à turma de yôga, por me ensinar e ser o meu momento de descontração quando eu mais precisava.

O tempo do doutoramento foi um tempo com um conteúdo rico e interessante ...

"Um conteúdo rico e interessante é, por outro lado, capaz de abreviar a hora e até mesmo o dia; mas, considerado sob o ponto de vista do conjunto, confere amplitude, peso e solidez ao curso do tempo, de maneira que os anos ricos em acontecimentos passam muito mais devagar do que aqueles outros, pobres, vazios, leves, que são varridos pelo vento e se vão voando."

Thomas Mann "A montanha mágica"

Obrigada a todos que fizeram parte desta aventura

ABSTRACT

The Western honey bee, *Apis mellifera* L., acts as a pollinator, thus playing a role in the ecosystem of paramount importance. However, the genetic integrity of many subspecies has been threatened by introgressive hybridization. In an attempt to reverse this trend, the main goals of this dissertation were (i) to reveal the genetic structure of one of the most complex and diverse subspecies in Europe, the Iberian honey bee (*Apis mellifera iberiensis*), and (ii) to develop molecular tools for the Iberian honey bee and its sister subspecies, the Dark honey bee (*Apis mellifera mellifera*), both belonging to the western European lineage M. These molecular tools can be employed for breeding and conservation programs in western and southern Europe.

The Dark honey bee *A. m. mellifera* has been severely threatened by hybridization with subspecies of eastern European (C lineage) ancestry, such as *A. m. ligustica* and *A. m. carnica*. Using 113 haploid honey bees collected from eight countries and genotyped with 1183 SNPs with the GoldenGate® Assay, five panels containing 48, 96, 144, 192 and 384 ancestry informative SNPs (fitted to the plexes of GoldenGate® Assays) were designed to estimate admixture proportions of C-lineage into *A. m. mellifera.* All SNP panels were able to estimate C-lineage admixture proportions highly concordant with those inferred from the 1183 SNP dataset ($r \geq 0.997$).

The discontinuation of Illumina's GoldenGate® Assay and the need of a standard method to examine the purity of honey bee populations in a wide geographical area were the motivation to design four multiplexed SNP assays to be genotyped using the iPLEX MassARRAY system, having as a baseline the 144-plex SNP panel. An accurate and cost-effective tool was provided with all genomic information for 117 SNPs for immediate application in genetic surveys and conservation management of *A. m. mellifera*. The performance of the assays was assessed against the data from 27 Whole-Genome (WG) sequences and using a set of individuals obtained from controlled crosses. In addition, sensitivity tests indicate that this genotyping system has the potential to detect C-lineage introgression at a low frequency (diluted as 1:20 in DNA pools or as 1 F1 hybrid: 7 *A. m. mellifera* individuals in tissue pools).

In contrast with *A. m. mellifera*, *A. m. iberiensis* populations exhibit a preserved complex genetic variation pattern. Although neutral processes have played an important role in shaping the Iberian diversity pattern, selection should not be ignored. Therefore, the WGS data of 87 Iberian honey bees were scanned for selection signals using three methods (Samβada, LFMM and PCAdapt) and two datasets (genomic and environmental). Candidate SNPs detected by at

least two methods were further examined using a haplotype-based method and protein modelling. Among the 830 SNPs exhibiting selection signals, 90.2% lie in non-coding regions, suggesting that regulatory changes are important in local adaptation. An enrichment of non-synonymous SNPs was also found, three of them leading to amino acid replacements, within or in the close vicinity of a functionally important site of proteins with functions related to lipids and transmembrane transport. Using both genetic and environmental data, candidate genes putatively under climate-driven adaptation were identified. Interestingly, membrane-related and circadian clock genes, which allow the organism to sense and fine-tune with environmental oscillations, are among the strongest candidate genes. This is particularly important in the context of rapid global change, helping to understand the mechanisms used by organisms to adapt to varying environmental conditions.

Likewise for *A. m. mellifera*, it is important for *A. m. iberiensis* to have a cost-effective molecular tool capable of accurately detecting C-derived introgression. Reduced assays of highly informative SNPs were developed from 176 WGs. In addition, both the effects of sample size and of sampling a geographically restricted area on the number of highly informative SNPs were tested. Results show that a bias is introduced when the sample size is small (N≤10) and when sampling only captures a fraction of a population's genetic diversity. The designed assays can be readily used for monitoring populations not only in the native range of *A. m. iberiensis* but also in the introduced range.

Another molecular marker widely used to assess the genetic diversity in honey bees is the mitochondrial intergenic $tRNA^{leu}$-cox2 region. Using mitogenome data from 123 individuals representing seven subspecies, three lineages (A, M and C) and three African sub-lineages (A_I, A_{II} and A_{III}), it was tested whether the information provided by this region is reliable for historical inference. While the mitogenome analysis supports the three evolutionary lineages defined by the $tRNA^{leu}$-cox2 intergenic region, it does not support the existence of the three African sub-lineages. Finally, different parts of the mitogenome provided distinct results, implying that the conclusions drawn from studies using only one locus need to be taken with caution.

Overall, in this dissertation a set of accurate and reliable tools was developed to be used in the preservation of the genetic integrity of honey bee populations of M-lineage ancestry. Moreover, new insights into genetic basis of Iberian honey bee local adaptation were provided by WGS and environmental data.

RESUMO

A abelha-europeia, *Apis mellifera* L., exerce um papel essencial no ecossistema como polinizador. No entanto, a sua integridade genética tem sido ameaçada por hibridação introgressiva. Na tentativa de contribuir para contrariar esta tendência esta dissertação tem os seguintes objetivos: (i) revelar a estrutura genética de uma das subespécies mais complexas da Europa, a abelha ibérica (*Apis mellifera iberiensis*) e (ii) desenvolver ferramentas moleculares para a abelha ibérica e para a abelha-negra (*Apis mellifera mellifera*), ambas pertencentes à linhagem Europeia Ocidental ou M. Estas ferramentas podem ser usadas tanto em programas de melhoramento como de conservação.

A abelha-negra tem sido ameaçada pela hibridação com subespécies da Europa Oriental (linhagem C), como a *A. m. ligustica* e a *A. m. carnica*. Usando como base 113 abelhas haploides provenientes de oito países genotipadas com 1183 na GoldenGate® Assay, foram desenvolvidos cinco painéis com 48, 96, 144, 192 e 384 SNPs (adequados para a tecnologia GoldenGate® Assay) com o objetivo de estimar introgressão da linhagem C em *A. m. mellifera.* Todos os painéis tiveram estimativas de introgressão semelhantes à obtida quando os 1183 SNPs foram usados (r≥0,997).

A descontinuação da GoldenGate® Assay e a necessidade de um método para monitorizar a introgressão numa vasta área geográfica motivaram o desenvolvimento de quatro painéis de SNPs apropriados para o sistema iPLEX MassARRAY, tendo como base o painel de 144 SNPs, anteriormente desenvolvido. No final foi divulgada uma ferramenta precisa e económica, juntamente com toda a informação genética dos 117 SNPs, para uma aplicação imediata em estudos genéticos e de conservação da *A. m. mellifera*. O desempenho deste painel foi avaliado por intermédio de 27 genomas e usando indivíduos obtidos por cruzamentos controlados. Adicionalmente foram efetuados testes de sensibilidade que demonstraram que esta técnica permite a deteção de alelos da linhagem C, mesmo quando em baixas frequências (diluído de 1:20 numa *pool* de DNA e na proporção 1 F1:7 *A. m. mellifera* numa *pool* de tecidos).

Contrariamente à *A. m. mellifera*, as populações de *A. m. iberiensis* ainda não estão ameaçadas e têm uma estrutura genética complexa. Os processos neutrais tiveram um papel importante na moldagem da diversidade genética, no entanto o papel da seleção não pode ser ignorado. Por isso, foram procurados sinais de seleção usando um total de 87 genomas de abelhas ibéricas, três métodos (Samβada, LFMM e PCAdapt) e dois tipos de dados (genómicos e

ambientais). Os SNPs detetados por pelo menos dois destes métodos foram analisados usando métodos haplótipicos e modelação de proteínas. Entre os 830 SNPs com sinais de seleção, 90,2% estão em regiões não codificantes, sugerindo que a regulação tem um papel crucial na adaptação local. Foi também encontrado um enriquecimento de SNPs não-sinónimos, três deles levam à substituição de aminoácidos situados dentro ou perto de locais funcionais de proteínas relacionadas com lípidos e transporte transmembranar. Usando dados genéticos ambientais foram identificados genes relacionados com adaptação a diferentes climas. Curiosamente, os candidatos mais fortes foram genes relacionados com a membrana e o relógio circadiano que permitem que o organismo detete e se adapte a variações ambientais. Estudos de adaptação local são especialmente importantes no contexto das mudanças climáticas, ajudando a perceber quais os mecanismos usados para a adaptação a diferentes condições ambientais.

Ferramentas moleculares económicas e precisas para estimar a introgressão da linhagem C são tão importantes para a *A. m. mellifera* como para a *A. m. iberiensis*, por isso quatro painéis ultra-reduzidos foram desenvolvidos usando o genoma completo de 176 indivíduos. Adicionalmente, os efeitos do tamanho da amostra e da amostragem geograficamente confinada foram avaliados no número de SNPs fixos. Verificou-se que existe um enviesamento quando o tamanho da amostra é ≤10 e quando a amostragem representa uma pequena porção da diversidade genética. Os painéis ultra-reduzidos podem ser utilizados na monitorização da integridade genética não só na área nativa mas também em locais onde a abelha ibérica foi introduzida.

Outro marcador muito usado na avaliação da diversidade genética das abelhas é a região intergénica $tRNA^{leu}$-cox2 do DNA mitocondrial. Usando dados do mitogenoma de 123 indivíduos de sete subespécies diferentes, três linhagens (A, M e C) e três sublinhagens africanas (A_I, A_{II} e A_{III}), foi testado se esta região é fidedigna para inferência histórica. As análises mitogenómicas suportam as três linhagens evolutivas definidas por essa região intergénica mas não sustentam a existência das três sub-linhagens africanas. É também de salientar que diferentes partes do mitogenoma fornecem diferentes resultados, sugerindo que as conclusões retiradas de estudos que utilizem um só locus devem ser tomadas com precaução.

Nesta dissertação foram desenvolvidas ferramentas precisas e fidedignas que podem ser usadas para a preservação da integridade genética das populações europeias de abelhas da linhagem M. Os dados de sequenciação e ambientais ajudaram na compreensão da base genética da adaptação local da abelha ibérica.

ABSTRAKT

Den västra honungsbiet, *Apis mellifera L.*, spelar en nyckelroll som pollinatör och tillhandahåller viktiga ekosystemtjänster. Den genetiska integriten hos många underarter hotats dock av introgressiv hybridisering. I ett försök att vända denna trend har de huvudsakliga målen för denna avhandling varit att (i) kartlägga den genetiska strukturen hos det iberiska biet (*Apis mellifera iberiensis*), en av de mest komplexa och varierande underarterna i Europa och att (ii) utveckla molekylära verktyg för den iberiska honungsbiet och dess systerunderart, den nordiska biet (*Apis mellifera mellifera*), som båda tillhör den västeuropeiska linjen M. Dessa molekylära verktyg kan användas för att föda upp och bevara det viktiga genetiska arvet i västra och södra Europa.

Det nordiska biet *A. m. mellifera* har varit allvarligt hotat av hybridisering med underarter av östeuropeisk härkomst (C-linjen), såsom *A. m. ligustic*a och *A. m. carnica*. Med hjälp av 113 haploida honungsbin som samlats in från åtta länder och genotypats för 1183 SNPar på en GoldenGate® Assay så konstruerades fem paneler innehållande 48, 96, 144, 192 och 384 ursprungsinformativa SNPar (anpassade till plexen för GoldenGate® Assays) för att uppskatta omfattningen av C-linjens tillblandning (admixture) i *A. m. mellifera*. Hos samtliga SNP-paneler uppskattades tillblandningsförhållanden för C-linjen som stämde väl överens med de som beräknades från 1183 SNP-datasetet ($r \geq 0{,}997$).

Avvecklingen av Illuminas GoldenGate® Assay och behovet av en standardmetod för att undersöka renheten hos honungsbi-populationer i ett brett geografiskt område blev motiven bakom utformningen av fyra multiplexade SNP-test som genotypades med hjälp av iPLEX MassARRAY-systemet, med en 144-plex SNP-panel som baslinje. Ett tillförlitligt och kostnadseffektivt verktyg togs fram för kartläggning av genomisk information i 117 SNPar och som redan nu kan användas vid genetiska undersökningar och bevarandearbeten av *A. m. mellifera*. SNP-testernas prestanda utvärderades gentemot helgenom-sekvensering (WGS) och en uppsättning individer erhållna från kontrollerade korsningar. Sensitivitetstest indikerade dessutom att detta genotypningssystem har potential att upptäcka C-linjens introgression vid låg frekvens (utspädd som 1:20 i DNA-pooler eller som 1 F1-hybrid: 7 *A. m. mellifera* individer i vävnadspooler).

I motsats till *A. m. mellifera*, uppvisar *A. m. iberiensis*-populationer ett bevarat och komplext genetiskt variationsmönster. Även om neutrala processer har spelat en viktig roll för att forma det iberiska mångfalden, är naturligt urval en process som inte kan ignoreras. Därför skannades WGS-data från 87 iberiska honungsbin efter tecken på selektion med hjälp av tre metoder (Samβada, LFMM och PCAdapt) och två dataset (genomiskt och miljömässigt).

Kandidat-SNPar som detekterades med åtminstone två metoder undersöktes vidare med hjälp av en haplotypbaserad metod och proteinmodellering. Bland de 830 SNPar som uppvisar selektionssignaler ligger 90,2% i icke-kodande regioner, vilket tyder på att förändringar i genreglering är viktiga vid lokal anpassning. En anrikning av icke-synonyma SNPar hittades också, tre av vilka som leder till aminosyrautbyten inom eller i närheten av funktionellt viktig delar hos proteiner med funktioner relaterade till lipid- och membrantransport. Genom att använda både genetiska och miljömässiga data identifierades kandidatgener för klimatdrivna anpassningar. Membranrelaterade och cirkadiska klockgener, som gör det möjligt för organismer att känna av och justera sig efter miljösvängningar, är intressant nog bland de starkaste kandidatgenerna. Detta hjälper oss förstå de mekanismer som organismer använder för att anpassa sig till olika miljöförhållanden, vilket är särskilt viktigt i en tid av snabba globala förändringar.

Precis som för *A. m. mellifera*, det är viktigt att ha ett kostnadseffektivt molekylärt verktyg som kan exakt detektera C-härledd introgression även för *A m. iberiensis*. Reducerade analyser av höginformativa SNPar togs fram från 176 helgenom. Dessutom testades vilka effekter provstorlek och provtagning av ett geografiskt begränsat område har på antalet höginformativa SNPar. Resultaten visar att en systematisk avvikelse introduceras när provstorleken är liten (N≤10) och när provtagningen endast fångar en bråkdel av en populations genetiska mångfald. Analyserna som utformats är redo att användas för övervakning av populationer, inte bara i det iberiska biets naturliga utbredningsområde, utan även i introducerade områden.

En annan molekylär markör som i stor utsträckning används för att bedöma den genetiska mångfalden hos honungsbina är den mitokondriella intergena $tRNA^{leu}$-cox2-regionen. Med hjälp av mitogenomdata från 123 individer som representerar sju underarter från tre linjer (A, M och C) och tre afrikanska undergrupper (*AI*, *AII* och *AIII*) så testades det om huruvida informationen från denna region är pålitlig för historisk analys. Trots att mitogenomsanalysen stöder de tre evolutionära linjerna definierade av $tRNA^{leu}$-cox2-intergena regionen så stöder den inte uppdelningen av de tre afrikanska undergrupperna. Olika delar av mitogenomen gav olika resultat, vilket innebär att slutsatserna från studier som endast använder ett lokus måste hanteras med viss försiktighet.

Sammantaget utvecklades det i denna avhandling exakta och tillförlitliga verktyg som kan användas för att bevara den genetiska integriteten hos M-linjens honungsbipopulationer. Med hjälp av helgenomsekvensering och miljödata tillkom nya insikter i den genetiska bakgrunden till lokal anpassning hos det iberiska honungsbiet.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS AND ACRONYMS

A
ABC Transporter: ATP binding cassette transporter
ABGD: Automatic Barcode Gap Discovery
ahb: Initials to assign Africanized honey bee genome (traces)-derived SNPs
AIMs: Ancestry-Informative Markers
Ala or A: Alanine
AMB: Initials to assign honey bee reference genome-derived SNPs
Amel_4.5: *Apis mellifera* genome assembly
AMOVA: Analysis of Molecular Variance
Arg or R: Arginine
Asn or N: Asparagine
Asp or D: Aspartic acid
AT: Atlantic Transect
ATP: Adenosine Tri-phosphate
ATP6: Adenosine Tri-Phosphate synthase subunit 6
ATP8: Adenosine Tri-Phosphate synthase subunit 8

B
BEEBASE: Bee genomic database
BIC: Bayesian Information Criterion
bp: Base pairs
bPTP: Bayesian Poisson Tree Processes
BWA: Burrows-Wheeler Aligner
BYM: Convolution Gaussian prior spatial mixture

C
CI: Confidence Interval
CIPRES: Cyberinfrastructure for Phylogenetic Research
Cld: Cloud cover
CLUMPAK: Clustering Markov Packager Across K
cM: Centimorgan
CoA: Coenzyme A
COX1 gene or **COI gene:** Cytochrome Oxidase subunit I
COX2 gene or **COII gene:** Cytochrome Oxidase subunit II
CpG: Regions of DNA where a cytosine occurs next to a guanine
CT: Central Transect
CTAB: Cetyltrimethylammonium-bromide
CV: Cross-validation
Cys or C: Cysteine
CYTB: Cytochrome b

D
DAVID: Database for Annotation, Visualization and Integrated Discovery
DNA: Deoxyribonucleic Acid

E

EHH: Extended Haplotype Homozygosity
est: Initials to assign honey bee expression sequence tag-derived SNPs
ESTs: Expressed Sequence Tags

F

FDR: False Discovery Rate
FLYBASE: Fruit fly genomic database
FreeBayes: Bayesian genetic variant detector
F_{ST}: Fixation index of genetic differentiation

G

G: Likelihood ratio
GATK: Genome Analysis Toolkit
GEA:Genetic-Environment Association
GenBank: Genetic sequence database
GIS: Geographic Information System
Gln or Q: Glutamine
Gly or G: Glycine
GO: Gene Ontology
GPS: Global Positioning System

H

H: Haplotype diversity
Hd: Number of haplotypes
His or H: Histidine
HiSeq: HiSeq Sequencing Systems

I

Iberian honey bee: *Apis mellifera iberiensis*
iHS: Integrated haplotype score
Ile or I: Isoleucine
In: Informativeness
Indels: Insertion or deletion of bases in the DNA
Ins: Insolation

J

JC: Jukes-Cantor

K

K: Number of cluster
K^+: potassium
K2P: Kimura-2-Parameter
Kcal: Kilocalorie
KJ: Kilojoules
Km: Kilometer
Km2: Square kilometer

L

Lat: Latitude
LD: Linkage Disequilibrium
Leu or L: Leucine
LFMM: Latent Factor Mixed Models
LG: Linkage Groups
Lineage A: African lineage
Lineage C: Eastern Europe lineage
Lineage M: Western and Northern European lineage
Lineage O: Near East and Asia lineage
LLS: Log-Likelihood Scores
Long: Longitude
l-rRNA: large ribosomal RNA
LYS or L: Lysine

M

m^2: Square meter
MAF: Minor Allele Frequency
MALDI-TOF: Matrix-assisted Laser Desorption Ionization-Time-of-Flight Mass Spectrometry
Max: Maximum
Mb: Megabase
MCL: Markov Cluster
MCMC: Markov Chain Monte Carlo
MEGA: Molecular Evolutionary Genetics Analysis
Met or M: Methionine
Mitotype: Mitochondrial haplotype
Mol: mole
MT: Mediterranean Transect
mtDNA: mitochondrial DNA

N

N_a: Mean number of alleles per locus
NCBI: National Center for Biotechnology Information
ND1: Nicotinamide adenine dinucleotide dehydrogenase subunit 1
ND2: Nicotinamide adenine dinucleotide dehydrogenase subunit 2
ND3: Nicotinamide adenine dinucleotide dehydrogenase subunit 3
ND4: Nicotinamide adenine dinucleotide dehydrogenase subunit 4
ND4L: Nicotinamide adenine dinucleotide dehydrogenase subunit chain 4L
ND5: Nicotinamide adenine dinucleotide dehydrogenase subunit 5
ND6: Nicotinamide adenine dinucleotide dehydrogenase subunit 6
N_e: Effective number of alleles
ng/µl: Nanogram/microlitre
NJ: Neighbor-Joining
nm: Nanometer
N_p: Number of private alleles

O

ONCOR: *a computer program for Genetic Stock Identification*
OrthoDB: Hierarchical catalog of animal, fungal and bacterial orthologs

P
PC: Principal Component
PCA: Principal Component Analysis
PCR - RFLP: Polymerase Chain Reaction - Restriction Fragment Length Polymorphism
PCR: Polymerase Chain Reaction
PDB: Protein Data Bank.
***Per*:** Period
Phe or F: Phenylalanine
Phyre2: Protein Homology/analogy Recognition Engine V 2.0
PLINK: Whole genome data analysis toolset
Prec: Precipitation
Pro or P: Proline
***P*-value:** It is a statistical probability of Karl Pearson

Q
Q. Matrix of membership proportions

R
***r*:** Pearson's correlation coefficient
RFLP: Restriction Fragment Length
Rh: Relative humidity
RMSD: Root-Median-Square Deviations
RNA: Ribonucleic Acid

S
S: Serine
SD: Standard Deviation
Ser: Serine
SnpEff: Genomic variant annotations and functional prediction toolbox
SNPs: Single Nucleotide Polymorphisms
srRNA: Small ribosomal RNA
SSR: Simple Sequence Repeat
STR: Short Tandem Repeat

T
Thr or T: Threonine
Tmax: Maximum temperature
Tmean: Mean temperature
Tmin: Minimum temperature
tRNA: transfer RNA
***tRNA*^{leu} :** *transfer RNA leucine*
Tyr or Y: Tyrosine

U
UGT: UDP-Glycosyltransferases
***uh*:** Unbiased haploid genetic diversity
UK: United Kingdom
US: United States
UTRs: Untraslated Regions

V
Val or V: Valine
VNTR: Variable Number Tandem Repeat

W
WG: Whole Genome
WGS: Whole Genome Sequencing

ΔΔG :Gibbs-free energy
π: nucleotide diversity
πXY: The average number of pairwise differences between populations
''': Haplotype with five Q elements
'': Haplotype with four Q elements
': Haplotype with three Q elements
|r|: Absolute correlation value
3′ UTR: 3′ Untranslated Region
5′ UTR: 5′ Untranslated Region
Å: Ångström

°C: Celsius degree

LIST OF TABLES

LIST OF FIGURES

Publications in the scope of the thesis

Papers published in international peer-review journals

- Muñoz I, **Henriques D,** Johnston JS, Chávez-Galarza J, Kryger P, Pinto MA (2015). Reduced SNP panels for genetic identification and introgression analysis in the Dark honey bee (*Apis mellifera mellifera*). PLoS ONE, 10(4): e0124365.
  DOI: 10.1371/journal.pone.0124365
- Impact factor 2015: 3.23

Papers submitted in international peer-review journals

- **Henriques D**, Browne KA, Barnett MW, Parejo M, Kryger P, Freeman TC, Muñoz I, Garnery L, Highet F, Jonhston JS, McCormack GP, Pinto MA. High sample throughput genotyping for estimating C-lineage introgression in the Dark honeybee: an accurate and cost-effective SNP-based tool. *Scientific Reports*.
  Submission date: 18th January 2018
  Impact factor 2016: 4.259
- **Henriques D**, Parejo M, Vignal A, Wragg D, Wallberg A, Webster, M, Pinto MA. Developing reduced SNP assays from whole-genome sequence data to estimate introgression in an organism with complex genetic patterns, the Iberian honeybee (*Apis mellifera iberiensis*). *Evolutionary Applications*
  Submission date: 6th January 2018
  Impact factor 2016: 5.671
- **Henriques D**, Wallberg A, Chávez-Galarza J, Johnston JS, Webster MT, Pinto MA. Local adaptation in Iberian honeybees: insights from whole genomes.
  To be submitted in *Scientific Reports*
  Impact factor 2016: 4.259

Papers in preparation

- **Henriques D**, Wallberg A, Chávez-Galarza J, Johnston JS, Webster MT, Pinto MA. Local adaptation in Iberian honeybees: insights from whole genomes.
  To be submitted in *Scientific Reports*
  Impact factor 2016: 4.259

- **Henriques D**, Chávez-Galarza J, Neves C., Quaresma A., Costa F., Rufino J, Pinto MA. Mitochondrial DNA patterns in honey bees: from mitogenomes to the popular intergenic tRNAleu-Cox2 region

Other papers published in international peer-review journal

- Chávez-Galarza J, Garnery L, **Henriques D**, Neves C. J, Loucif-Ayad W, Jonhston JS, Pinto MA 2017. Maternal variation of *Apis mellifera iberiensis*: further insights from a large scale study using sequence data of the tRNAleu-cox2 mitochondrial intergenic region. Apidology: 48: 533-544. DOI: 10.1007/s13592-017-0498-2.
- Muñoz I, **Henriques D**, Jara L, Johnston JS, Chávez-Galarza J, De La Rúa P, Pinto MA 2017. SNPs selected by information content outperform randomly selected microsatellite loci for delineating genetic identification and introgression in the endangered dark European honeybee (*Apis mellifera mellifera*). Mol Ecol Resour 17: 783-795. doi: 10.1111/1755-0998.12637
- Parejo M, Wragg D, **Henriques D**, Vignal A, Neuditschko M 2017. Genome-wide scans between two honeybee populations reveal putative signatures of human-mediated selection. Animal genetics 48: 704-707. doi: 10.1111/age.12599
- Muñoz I, **Henriques D**, Johnston JS, Chávez-Galarza J, Kryger P, Pinto MA 2015. Reduced SNP panels for genetic identification and introgression analysis in the dark honey bee (*Apis mellifera mellifera*). PLoS One 10: e0124365. doi: 10.1371/journal.pone.0124365
- Chávez-Galarza J, **Henriques D**, Johnston JS, Carneiro M, Rufino J, Patton JC, Pinto MA 2015. Revisiting the Iberian honey bee (*Apis mellifera iberien*sis) contact zone: maternal and genome-wide nuclear variations provide support for secondary contact from historical refugia. Molecular Ecology 24: 2973-2992. doi: 10.1111/mec.13223
- Pinto MA, **Henriques D**, Chávez-Galarza J, Kryger P, Garnery L, van der Zee R, Dahle B, Soland-Reckeweg G, De la Rúa P, Dall' Olio R, Carreck NL, Johnston JS 2014. Genetic integrity of the Dark European honey bee (*Apis mellifera mellifera*) from protected populations: a genome-wide assessment using SNPs and mtDNA sequence data. Journal of Apicultural Research 53: 269-278. doi: 10.3896/ibra.1.53.2.08

Other papers submitted in international peer-review journals

- Parejo M, **Henriques D**, Pinto MA, Soland-Reckewege G, Neuditschkoa M 2018. Empirical comparison of microsatellite and SNP markers to estimate introgression in *Apis mellifera mellifera*. Journal of Apicultural Research submitted.
- Pérez-Rodriguez F, Neves CJ, **Henriques D**, Pinto MA. A note to transfer a generic database architecture for storing chronological data from research in apiaries. Journal of Apicultural Research
-

Chapter I.

Motivation, objectives and thesis outlines

Motivation

The Western honey bee (*Apis mellifera* L.) is one of the most important managed pollinators and is currently facing a growing number of threats, largely influenced by the modern beekeeping activities. In Western Europe, one such threat is the large-scale introduction of commercial strains and foreign subspecies, mainly *A. m. ligustica* and *A. m. carnica* (both from C-lineage). Despite its extensive native distribution, the genetic integrity of *A. m. mellifera* (M-lineage), known as the Dark honey bee, is now severely compromised by introgressive hybridization to such an extent that in some places it has virtually been driven to extinction. The knowledge that reduced adapted genetic diversity threatens both managed and unmanaged populations led to numerous initiatives to protect and bring back the endangered Dark honey bee. The success of these initiatives relies on molecular tools capable of accurately detecting varying levels of C-derived introgression in a time- and cost-effective manner. For that reason, the first goal of this dissertation was to develop a tool to estimate the C-lineage introgression into *A. m. mellifera*.

Conversely to *A. m. mellifera*, the Iberian honey bee *A. m. iberiensis* exhibits a preserved complex genetic variation pattern. While neutral processes have played an important role in shaping the Iberian honey bee diversity pattern, selection is a force that cannot be ignored. Therefore, the second goal of this thesis was to scan the Iberian honey bee genome for selection signals employing an integrative approach of environmental data and whole-genome (WG) resequencing data. The findings of the selection scan are important to understand the complexity of this subspecies and the unique genetic background that enables its adaptation to the different climates in Iberia.

Until now, the native range of *A. m. iberiensis* has not been threatened by hybridization, although this scenario might change as many young beekeepers are attracted by the advertised prolificity and gentleness of commercial C-lineage strains. Conservation measures should be applied before unique combinations of traits shaped by natural selection are lost. As for *A. m. mellifera*, it is important to have a molecular tool capable of accurately detecting variation levels of C-derived introgression in a time- and cost-effective manner for *A. m. iberiensis*. Whilst high-throughput SNP analysis on WG sequence data may be required in selection and demographic studies, it is not affordable for conservation management applications. So, the third goal of this thesis was to employ WG sequence data to develop low-density and robust SNP-panels, capable of

accurately detecting varying levels of C-derived introgression in a time- and cost-effective manner to be applied in *A. m. mellifera* and *A. m. iberiensis* populations.

Another molecular marker widely used in conservation centres is the mitochondrial intergenic $tRNA^{leu}$-cox2 region. Given differential mutation rates in different partitions of the mitochondrial genome, the patterns of the mitogenome, individual coding-regions and the intergenic $tRNA^{leu}$-cox2 region were compared to understand if they are concordant and if the $tRNA^{leu}$-cox2 region is reliable for historical inference. To achieve this final goal, WG sequencing data mapped in the mitochondrial DNA was used.

I hope the findings of this PhD work will be helpful in designing breeding and management programs in honey bee populations of northern and western Europe.

Objectives

- To develop time- and cost-effective molecular tools, capable of accurately detecting varying levels of C-derived introgression into European M-lineage populations (Chapters III, IV and VI).
- To identify signatures of selection in Iberian honey bees to understand the molecular basis of adaptation (Chapter V).
- To assess concordance of different mitochondrial regions and whether the $tRNA^{leu}$-cox2 region is reliable for historical inference (Chapter VII).

Thesis outline

Chapter I

Chapter I describes the context, motivation and goals of this thesis, as well as its global structure.

Chapter II

Chapter II provides an overview of the state of the art related with the topics of this thesis, describing briefly the genus *Apis* and the evolutionary lineages, the current threats faced by honey bees, with a special focus on the problem of hybridization, and the main molecular markers used to study honey bee diversity.

Chapter III

In this chapter, reduced panels of ancestry-informative markers (AIMs) were selected from 1183 single-nucleotide polymorphisms (SNPs) with the aim of providing accurate estimates of the level of C-lineage introgression into *A. m. mellifera*.

Chapter IV

Using as a baseline the 144-SNP panel constructed in chapter III, four multiplexed assays were designed to be genotyped using the iPLEX MassARRAY system of Agena BioScience™, which is amongst the most cost-effective platforms for medium SNP throughput and high sample throughput genotyping. The four customized SNP assays were fully tested and validated by comparing with WG sequence data. In addition, the sensitivity of the assays was tested in pools of tissue and DNA.

Chapter V

In Chapter V, a combination of outlier and genetic-environment association (GEA) methods were employed to identify signatures of selection in the Iberian honey bee using 87 WG. Genes putatively involved in adaptation to different environments as well as environmental factors that might act as selective pressures in Iberian honey bees were identified. The findings will be helpful in designing breeding and management programs for Iberian honey bees.

Chapter VI

In Chapter VI, cost-effective and robust reduced SNP assays were constructed to infer C-lineage introgression into *A. m. iberiensis* using as a baseline 176 WGS. Taking advantage of the large and comprehensive WG dataset, the effect of sample size and sampling a geographically restricted area on the development of reduced SNP assays was tested. The reduced SNP assays were validated in holdout and simulated sets.

Chapter VII

Mitochondrial DNA (mtDNA), more specifically the $tRNA^{leu}$-cox2 region, has been the marker of choice in honey bee phylogeographical studies and in assessing the conservation status of protected populations. Using a large mtDNA-WG dataset, the phylogenetic congruence between the mitogenome, the $tRNA^{leu}$-cox2 intergenic region and different individual mitochondrial genes was investigated.

Chapter VIII

In this chapter, the overall conclusions are presented discussing the relevance of this work. To prompt further developments, ideas for future studies are suggested.

Chapter IX

The last chapter includes supplementary data not shown in the other chapters and the original version of the chapters that have been published or submitted. Some supplementary data is present in digital format.

Chapter II.

General introduction

Evolutionary history

The genus *Apis* belongs to the Apidae family, characterized by the pollen basket, the centre of origin of this genus is not consensual, while Ruttner (1988) described the centre of origin in Asia, the region with most number of living species, recently Kotthoff *et al.* (2013) using fossil data have suggested an apparent European origin. This genus contains 12 species forming three groups: cavity-nesting bees (*Apis mellifera, A. cerana, A. koschevnikovi, A. nuluensis, A. breviligula*), giant bees (*A. dorsata, A. laboriosa, A. binghami, A. nigrocincta, A. indica*), and dwarf bees (*A. florea, A. andreniformis*) (Arias & Sheppard 2005; Raffiudin & Crozier 2007).

Figure II-1 - Map with the geographical distribution of the 31 subspecies of *A. mellifera*. The colour of the names indicates the evolutionary lineage: A-African lineage (red), M-Western and Northern European linage (blue), C- Eastern European lineage (orange) and O- Middle East and Western Asia lineage (green). *A. sossimai*, *A. m. taurica* and *A. m. artemisia* have little information on their evolutionary lineage (black).

Currently, because of its economic relevance, the Western honey bee (*Apis mellifer*a) is the most widely distributed bee species around the World. However, its natural distribution encompasses Middle East, western Asia, Europe and Africa where it diversified into 31 subspecies (Chen *et al.*, 2016; Engel, 1999; Meixner *et al.*, 2011; Smith & Glenn, 1995) (Figure II-1).

Phylogeographical studies using a number of different genetic markers (Cornuet & Garnery, 1991; Ruttner, 1988; Whitfield *et al.*, 2006) grouped this wide-ranging diversity into four major lineages: African (A), Middle Eastern (O), Southeastern European (C), and Western and Northern European (M) (Figure II-2).

From the 31 currently recognized subspecies, 10 are endemic to Europe and belong to the M and C evolutionary lineages. In Europe, lineage M includes only two subspecies: the Dark honey bee *Apis mellifera mellifera* and the Iberian honey bee *Apis mellifera iberiensis*. Yet, these subspecies cover the largest territory in Europe with *A. m. iberiensis* occupying the Iberian Peninsula and *A. m. mellifera* ranging from France in the south to Scandinavia in the north, and from Ireland and the United Kingdom in the west to the Ural Mountains in the east (Ruttner, 1988). Lineage C occurs in a smaller geographical area confined to the Apennine and Balkan peninsulas and includes the most widely kept honey bee subspecies: the Italian *Apis mellifera ligustica* and the Carniolan *Apis mellifera carnica*.

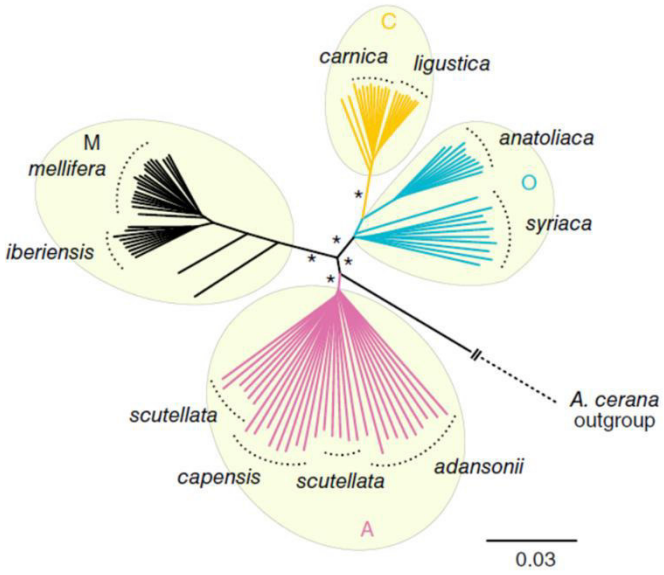


Figure II-2 - Neighbour-joining tree showing the grouping of 10 subspecies into four different evolutionary lineages (Wallberg *et al.*, 2014).

While *A. m. mellifera* patterns have been consensually explained by post-glacial re-colonization from an Iberian refuge, interpretation of *A. m. iberiensis* variation has proved more challenging. Iberia possesses a high physiographic complexity, with several large mountain ranges, and due to its geographical position, is under the influence of both the North Atlantic and the Mediterranean Sea. These features have shaped a diverse array of climates (including desert,

Mediterranean, Alpine, and Atlantic) and plant communities with variable flowering peaks to which the Iberian honey bee had to adapt.

The numerous Iberian phylogeographical studies revealed highly complex and incongruent diversity patterns, which led to two competing hypotheses for the M-lineage origin (Figure II-3). Early surveys of morphology (Ruttner, 1988) and allozymes (Smith & Glenn, 1995) revealed a gradual cline extending from Africa to northern Europe, with Iberian honey bees showing intermediate phenotypes. This pattern raised a hypothesis of primary intergradation for M-lineage origin (Ruttner, 1988; Smith & Glenn, 1995). However, the northeastern-southwestern Iberian cline formed by two highly divergent A and M maternal lineages was more compatible with a secondary intergradation hypothesis (Franck *et al.*, 1998; Smith *et al.*, 1991) (Figure II-3). To complicate matters further, microsatellites support neither hypothesis. Microsatellites showed no sub-division in Iberia, virtually no differentiation between *A. m. iberiensis* and *A. m. mellifera*, and a sharp disruption between M and A lineages (Franck *et al.*, 1998).

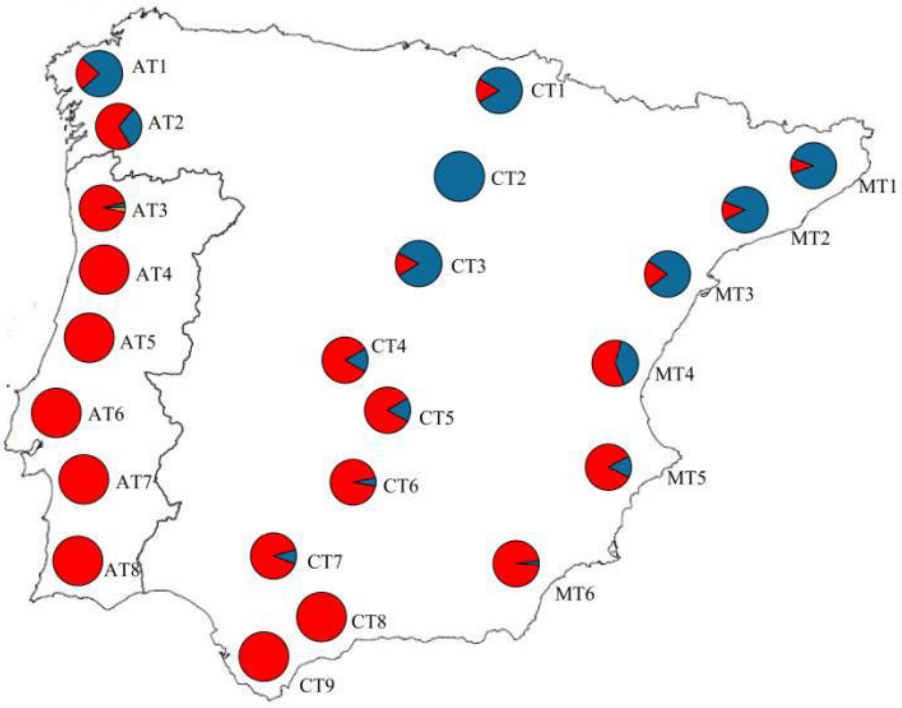


Figure II-3 - Maternal pattern in Iberian honey bees inferred from the intergenic $tRNA^{leu}$-cox2mtDNA region. The pie charts displayed on the Iberian map show the frequencies of the A (in red) and M (in blue) mitotypes at each sampling site (Chávez-Galarza *et al.*, 2015).

Ad hoc explanations have been proposed for the observed *A. m. iberiensis* patterns. For example, introgression of African haplotypes has been explained by contrasting views of recent

gene flow: one claiming human-mediated introduction of African colonies during Moorish invasions (Franck *et al.*, 1998) and the other claiming post-glacial expansion of northern African honey bees (Cánovas *et al.*, 2008). At the same time, selection has repeatedly been invoked to explain morphological (Ruttner, 1988), allozymic (Smith & Glenn, 1995), and mitochondrial (Franck *et al.*, 1998) clines.

General threats

The honey bee provides valuable pollination services, which play a critical role in ecosystem functioning and in food production for humanity. In spite of this, honey bees have been subject to a growing number of human-mediated threats. Over time beekeeping evolved from honey hunting to more sophisticated techniques, which also influenced the natural range of honey bees as well as their genetic composition (Crane, 1999). Colonies have been selected for their low defensive behaviour, low swarming, small propolis usage and high honey production. For that purpose, also different subspecies are crossed promoting the gene-flow and admixture between different populations with completely different genetic background, seriously compromising the genetic integrity of locally adapted ecotypes (De la Rúa *et al.*, 2009; Meixner, 2010; Pinto *et al.*, 2014). Another recurrent threat also related with large-scale commercial movements of honey bees worldwide is the risk of introduction and spread of new diseases (McMahon *et al.*, 2016; Mutinelli *et al.*, 2014) such as the ectoparasitic mite, *Varroa destructo*r, with its associated viruses and the microsporidia *Nosema cerenae* (Guzmán-Novoa *et al.*, 2010; Schäfer *et al.*, 2010). At this moment, the majority of European *Apis mellifera* colonies could not survive without control measures for *Varroa destructor* (Rosenkranz *et al.*, 2010). There are endless factors that could also have an harmful impact in the colony health including nutrition (vanEngelsdorp & Meixner, 2010), pesticides (Johnson *et al.*, 2010; Neumann & Blacquière, 2017) and climate change (Le Conte & Navajas, 2008). For instance, the increase of atmospheric carbon dioxide levels, a consequence of climate change, reduces pollen quality which is vital for honey bee larval development (Brodschneider & Crailsheim, 2010). Even the plant distribution shift has an impact since it affects the availability of resources important to the bees (Le Conte & Navajas, 2008).

Introgression (causes and consequences)

Introgression is the result of the flow of alleles from one species or subspecies to another. The role of introgression and admixture in conservation is a dilemma: While natural admixture may be an important evolutionary force in speciation and maintenance of genetic diversity (Dowling & Secor, 1997; Nolte & Tautz, 2010) admixture induced by human activities may contribute, either directly or indirectly, to the extinction of many taxa (Rhymer & Simberloff, 1996). Introduction of species, subspecies and habitat modifications has caused increased rates of admixture with native flora and fauna. Thereby, introgression can generate extinction and irretrievable loss of combinations of genotypes throughout the entire genome (Allendorf & Luikart, 2007).

The polyandric mating system of the honey bee together with the large-scale circulation of commercial queens and package honey bees promotes sympatry and gene flow. There are three intensively used *A. mellifera* subspecies, which are known for their docile nature and high productivity: *A. m. ligustica, A. m. carnica and A. m. caucasica*. The Italian honey bee (*A. m. ligustica*) was introduced in Northern Europe, America, Australia and the Canary Islands (De la Rúa *et al.*, 2009; Franck *et al.*, 2000). The Carniolan honey bee (*A. m. carnica*) was introduced in many regions of whole Europe (De la Rúa *et al.*, 2009; Parejo *et al.*, 2016). Lastly, the Caucasian honey bee (*A. m. caucasica*) has been intensively used by beekeepers through Western Turkey, Bulgaria, Russia, Ukraine, Germany and France (Ruttner, 1988). Currently, the artificial strain Buckfast is distributed around the world. This strain has been selected for superior honey production and lower defensive behaviour and it is mostly of C-derived ancestry. The human-mediated gene flow has changed the natural distribution of the honey bee to such an extent that the formerly widest-spread European subspecies, *A. m. mellifera*, is threatened by extinction through introgression and replacement from highly divergent commercial strains (De la Rúa *et al.*, 2009; Jensen *et al.*, 2005; Pinto *et al.*, 2014; Soland-Reckeweg *et al.*, 2009). Even when there is no complete replacement, the admixture between two divergent lineages reduces the frequency of locally adapted gene complexes, that have been shaped by natural selection over extended periods, finally, leading to an increased likelihood of reduced survival rates among native colonies (Allendorf & Luikart, 2007; De la Rúa *et al.*, 2013). While modern beekeeping led to the extinction of some unique gene complexes, it is important to recognize that the genetic diversity is crucial for a long-term sustainable beekeeping activity.

Tools and importance for preserving genetic diversity

Genetic diversity is required for populations and species to evolve in response to environmental change, such as climate and habitat changes. Large scale queen breeding, artificial selection and widespread propagation of selected stock reduce the effective population size and, consequently, lead to a loss of genetic diversity (Estoup *et al.*, 1995). Yet, maintaining locally adapted subspecies is crucial for the long-term sustainability of *A. mellifera* subspecies (De la Rúa *et al.*, 2013; vanEngelsdorp & Meixner, 2010). Reciprocal translocation experiments have shown that local honey bees have longer survivorship (Büchler *et al.*, 2014) and lower pathogen loads (Francis *et al.*, 2014) than the introduced subspecies. The recognition that native genetic diversity is fundamental for the bees that are facing the challenges of a rapidly changing world, led to the establishment of several conservation programs and protected areas throughout Europe (De la Rúa *et al.*, 2009). One of the earliest conservation programs enacted by law is that implemented by the Danish Beekeepers Association and the Læsø Beekeepers. Scottish government approved an order to protect the *A. m. mellifera* on the islands of Colonsay and Oronsay [The Bee Keeping (Colonsay and Oronsay) Order 2013]. Another European reserve was created in the United Kingdom. In addition to these, other *A. m. mellifera* conservation efforts, although not enacted by law, are underway in France, Netherlands, Norway, Switzerland, Ireland, and Belgium (see the website "http://www.sicamm.org" run by the International Association for the Protection of the European Dark bee). Other subspecies have also been protected, for instance *A. m. carnica* in Slovenia and Austria (De la Rúa *et al.*, 2009). While *A. m. carnica* is widely used in beekeeping activities only a small portion of its diversity is explored; Therefore, it is important maintain the original diversity of this subspecies.

The implementation of conservation programs across Europe in an attempt to recover and protect *A. m. mellifera* is very important, but other actions should also be considered towards preserving the genetic diversity of subspecies that are still not endangered, as it is the case of *A. m iberiensis*. In this way, it is possible to maximize the genetic diversity that can be conserved.

Across western and northern Europe different genetic tools have been applied to monitor *A. m. mellifera* populations with the aim of protecting and bringing back the endangered Dark honey bee. The success of these initiatives relies on molecular tools capable of accurately detecting varying levels of C-derived introgression in a time- and cost-effective manner. In several conservation programs, the honey bee breeding stock has been routinely identified through wing

morphometry. However, based on data from Africanized honey bees (Guzmán-Novoa *et al.*, 2010), wing morphometry is likely unable to detect low levels of C-lineage introgression into *A. m. mellifera*. The mitochondrial intergenic $tRNA^{leu}$-cox2 region has also been widely used, although it only represents the maternal variation, and thus important information may be missed (Bertrand *et al.*, 2015). The limitations of wing morphometry and the intergenic $tRNA^{leu}$-cox2 region are overcome by microsatellites that have been used more recently (Jensen *et al.*, 2005; Soland-Reckeweg *et al.*, 2009). While adoption of microsatellites represented a major step in conservation management of *A. m. mellifera* (Soland-Reckeweg *et al.* 2009), recent studies have shown that a reduced number of high-graded SNPs (Muñoz *et al.*, 2015) outperform the multiallelic marker in estimating introgression (Muñoz *et al.*, 2017; Parejo *et al.*, 2018).

Mitochondrial DNA

The honey bee mitochondrial genome (Figure II-4) is small (~16 Kb), circular, compact (very few duplications, no introns, and short intergenic regions), and it is a haploid chromosome. It contains 13 protein-coding genes (COX1-COX3, ND1-ND6, NDL4, CYTB, ATP6 and ATP8) that encode for subunits of the electron transport chain, which is involved in numerous processes such as energy production and heat generation. It also encodes 24 genes for the translation machinery, two of them are ribosomal (l-rRNA and s-rRNA), encoding the RNA components of the mitochondrial ribosome, and the other 22 genes are transfer (tRNAs) RNAs (Crozier & Crozier, 1993).

Since Avise and Ellis (1986) supported the usefulness of mitochondrial DNA (mtDNA) to study genetic variation, fragments or the entire molecule have been widely adopted in numerous animals, and the honey bee is no exception. The popularity of mtDNA marker in honey bees can be explained by practical reasons. Firstly, mtDNA has a maternal inheritance which determines that a single individual (worker or drone) per colony is required to establish the mitotype of the entire colony (Evans *et al.*, 2013). Secondly, the lack of recombination, rapid mutation rate, and (4X) smaller effective size, and, consequently, a shorter expected time to reciprocal monophyly between geographic region (Galtier *et al.*, 2009; Meixner *et al.*, 2013) make this marker suitable to infer recent events and the interpretation of data is straightforward.

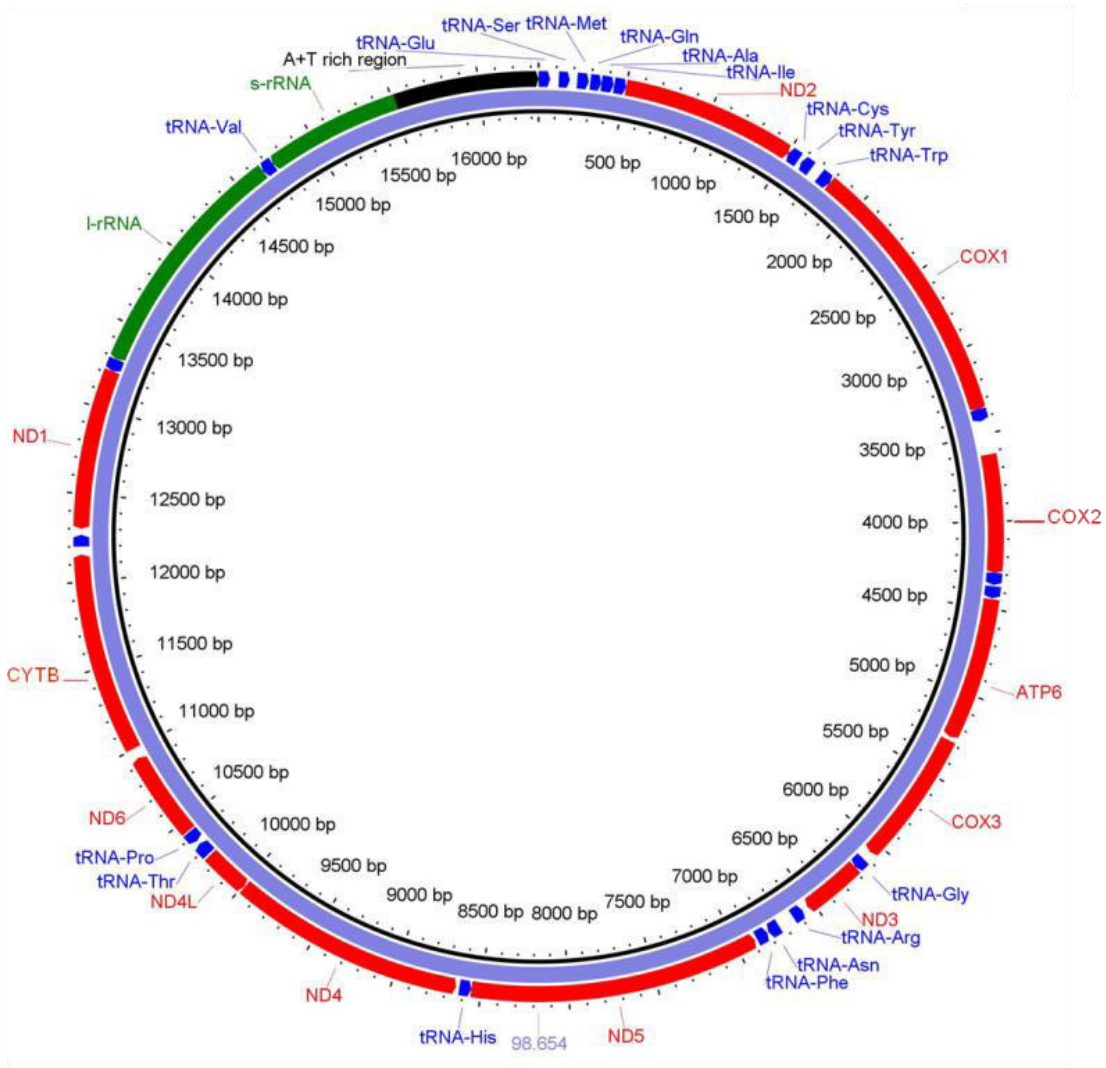


Figure II-4 - Reconstruction of the honey bee mitochondrial genome of *A. m. ligustica* using the MITObim (Hahn *et al.*, 2013). The tRNAs genes are denoted by short blue bars. The l-rRNA and s-rRNA are indicated with green bar. Protein-encoding genes are denoted by red bars, and the A+T rich.

Different arguments have been used to explain the mtDNA elevated mutation rate (5–10 times higher than nuclear DNA), such as the lack of histones, less efficient DNA repair mechanisms and a more mutagenic local environment (Lynch & Walsh, 2007). While mtDNA behaves as a single locus, the mutation rate is highly heterogenic along the genome, which suggests that the different parts of the mitogenome may lead to different results (Ballard & Whitlock, 2004; Keis *et al.*, 2013).

Different molecular methods have been applied to different parts of the mtDNA to assess maternal variation in honey bees. To have an overview of the genes that have been used to study genetic variation in honey bees, a data survey was performed using the Web of Science main

database. Articles published using mtDNA in honey bees (using the keywords "honey bee", "*Apis mellifera*", "mtDNA"and "mitochondrial") were searched for in this database in December 2017, resulting in 182 records. The first studies were performed after it was possible to purify the mitochondrial DNA from nuclear genomes using either ethidium bromide-CsCl gradients (Lansman *et al.*, 1981) or differential centrifugation on sucrose gradients (Cornuet & Garnery, 1991). The maternal variation was assessed using RFLP (Restriction Fragment Length Polymorphism) methods, in which the difference between two sequences is detected by the presence of fragments of different lengths after the digestion with restriction enzymes, in this case, applied to the entire molecule. The most common restriction enzymes used had a 4-base (eg: *Alu*I, *Hin*fI, *HPA*II, *RSA*I) and 6-base (eg: *Acc*I, *Ava*I, *Bcl*I, *Bgl*II, *EcoR*I, *Hin*dII, *Hind*II, *Hind*III, *Nde*I, *Pst*I, *Pvu*II, *Spe*I, *Xba*I, *Xbu*I) recognition site. Among this plethora of restriction enzymes, the *EcoR*I was used in 21 of the 25 articles that assessed the genetic variation using the entire mtDNA molecule. This RFLP method enabled to distinguish African from European subspecies and, within the European subspecies, between C-lineage (*A. m. carnica*, *A. m. ligustica*) and M-lineage (*A. m. mellifera*, *A. m. iberiensis*) subspecies (Schiff & Sheppard, 1996; Sheppard *et al.*, 1991). There are different studies that used the RFLP of the entire mtDNA, addressing both honey bee populations in the native range in Europe (Arias *et al.*, 2006; Garnery *et al.*, 1992; Sinacori *et al.*, 1998; Smith *et al.*, 1991) and Africa (Meixner, 2010), but also in the introduced range in the neotropics, mainly to study the Africanization process (Hall & Muralidharan, 1989; Schiff & Sheppard, 1993; Smith & Brown, 1988). While these RFLP surveys support the three morphological lineages (M, A and C) proposed by Ruttner (1988), and in some cases some subspecies (Meixner *et al.*, 2013), it is not diagnostic for most subspecies and it is very laborious. Another disadvantage is the requirement of large amounts of non-degraded DNA (Meixner *et al.*, 2013).

The description of the mitochondrial sequence of some insects (eg. *Drosophila yakuba*) (Clary & Wolstenholme, 1985) opened the possibility for using "universal" or "insect" primers (Cornuet & Garnery, 1991) for studying specific mtDNA fragments through polymerase chain reaction (PCR). Soon, the RFLP technique started to be replaced by the PCR-RFLP; here the restriction enzymes were also used but instead of the entire molecule just fragments were used. One advantage of this technique is the capacity of producing large amounts of DNA from a small amount of sample (Cornuet & Garnery, 1991). Finally, the sequencing of the mitochondrial genome of *A. m. ligustica* by Crozier and Crozier (1993) increased exponentially the number of

works that used the mtDNA (152 out of 161 studies were performed after 1993). Different regions and enzymes have been used depending of the aim and the subspecies (Table II-1).

Table II-1 - PCR-RFLP assays used to study genetic diversity in honey bees (Meixner *et al.*, 2013).

Mt fragments	Restriction Enzyme	Cleaved fragment	Uncleaved fragment	Authors
Cytochrome *b*	*Bgl* II	European	African	Crozier *et al.* (1991)
L-rRNA	*EcoR* I	*A. m. ligustica*, *A. m. carnica*, *A. m. caucásica*	*A. m. mellifera*, *A. m. iberiensis* of lineage M	Hall and Smith (1991)
COX1	*Hinc* II	*A. m. mellifera*, *A. m. iberiensis* of lineage M	*A. m. ligustica*, *A. m. carnica*, *A. m. caucasica*	Hall and Smith (1991)
	Xba I	*A. m. ligustica*, *A. m. carnica*, *A. m. caucásica*	*A. m. mellifera*, *A. m. iberiensis* of lineage M	Hall and Smith (1991)
	Hin fI	*A. m. lamarckii*	Non-*A. m. lamarckii*	Nielsen *et al.* (2000)
	Nco I	*A. m. macedonica*	*A. m. adami* *A. m. cecropia* *A. m. cypria*	Bouga *et al.* (2005) Stevanovic *et al.* (2010)
	Sty I	*A. m. macedonica*	*A. m. adami* *A. m. cecropia* *A. m. cypria*	Bouga *et al.* (2005) Stevanovic *et al.* (2010)
	Ssp I	Greek	Bulgarian	Ivanova (2010)
ND5	*Alu* I	*A. m. macedonica*	*A. m. adami* *A. m. cecropia* *A. m. cypria*	Bouga *et al.* (2005)
	Hinc II	Greek	Bulgarian	Ivanova (2010)
	Fok I	Greek	Bulgarian	Ivanova (2010)
$tRNA^{leu}$-COX2	*Dra* I			Garnery *et al.* (1993)

From the 13 protein-coding genes, eight have been used to study the genetic variation in the honey bee. In five of them (ND2, ND4, COX2, COX3 and ATP6) the variation has been assessed by sequencing. In the other three protein-coding genes (COX1, CYTB, ND5), the ribosomal gene l-rRNA and in the $tRNA^{leu}$-cox2 intergenic region the variation was examined using different restriction enzymes (Table II-1).

Some PCR-RFLP assays have been applied to distinguish very specific populations, such is the case of *A. m. macedonica* (Bouga *et al.*, 2005; Stevanovic *et al.*, 2010) and the Greek and Bulgarian honey bees (Ivanova, 2010). On the other hand, other assays have been applied to answer to specific problems. The CYTB/*Bgl*I assay has been important to study the Africanization process (Mortensen & Ellis, 2015; Pinto *et al.*, 2003; Pinto *et al.*, 2007) because it allows to

distinguish between European and African populations. For the Africanization studies other PCR-RFLP assays are required to distinguish the subspecies from north and south Africa and the different European populations (Figure II-5), which explains why the CYTB/*Bgl*II assay was mainly used in combination with other PCR-RFLP assays (Coulson *et al.*, 2005; Pinto *et al.*, 2004; Rangel *et al.*, 2016). The most common assays applied together with the CTYB/*Bgl*II were the l-rRNA/*EcoR*I (11 times), the COX1/*Hind*II and the COX1/*Xba*I (five times each), which distinguish the Eastern from the Western European populations (Figure II-5).

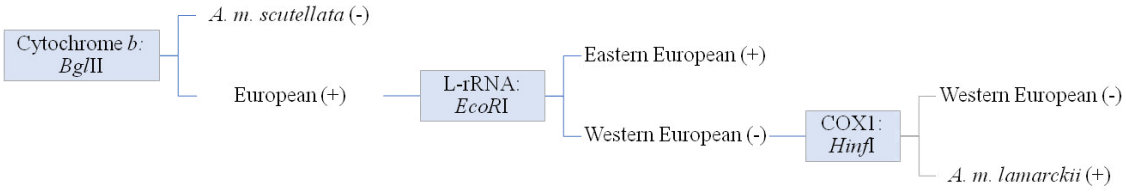


Figure II-5 - The PCR-RFLP assay used to study the Africanization process based on Pinto et al. (2004). The "+"sign indicates restriction site, and "-" indicates its absence.

Among the different PCR-RFLP assays described in Table II-1, the most popular is the tRNAleu-cox2, commonly known as *Dra*I test (102 out of 182 articles used this assay, Figure II-6). It consists of the PCR amplification of the intergenic region located between the $tRNA^{leu}$ and COX2 genes (originally named COI-COII intergenic region) followed by digestion with the *Dra*I restriction enzyme (Garnery *et al.*, 1993). While traditionally this is a PCR-RFLP assay, some studies sequenced the intergenic region to obtain a greater resolution (Collet *et al.*, 2006; Franck *et al.*, 2001; Kasangaki *et al.*, 2017; Pinto *et al.*, 2014; Pinto *et al.*, 2012; Shaibi *et al.*, 2009; Techer *et al.*, 2017; Techer *et al.*, 2015). The intergenic region has been used in a wide range of native (Garnery *et al.*, 1998; Gruber *et al.*, 2013; Meixner *et al.*, 2013; Pinto *et al.*, 2014; Pinto *et al.*, 2013) and introduced (Clarke *et al.*, 2002; Prada *et al.*, 2009; Sheppard *et al.*, 1999) populations and, due to the high information content, it allowed to shed light on the phylogeography of some subspecies (De la Rúa *et al.*, 2006; De la Rúa *et al.*, 1998; Franck *et al.*, 2001). From a practical point of view, it was also very useful to detect C-lineage introgression into *A. m. mellifera*, mainly in conservation areas (Jensen & Pedersen, 2005b; Pinto *et al.*, 2014) and to study the Africanization process in the neotropics (Clarke *et al.*, 2002). The $tRNA^{leu}$-cox2 is highly informative because it combines size and sequence variation. However, this region does not identify honey bees at the subspecies level, which is particularly problematic for C-lineage subspecies. For instance, Stevanovic *et al.* (2010) showed that it is necessary to use COX1 gene segments digested with

restriction enzymes *Nco*I and *Sty*I in order to discriminate *A. m. macedonica* from *A. m. carnica* which have the same C2d mtDNA haplotype.

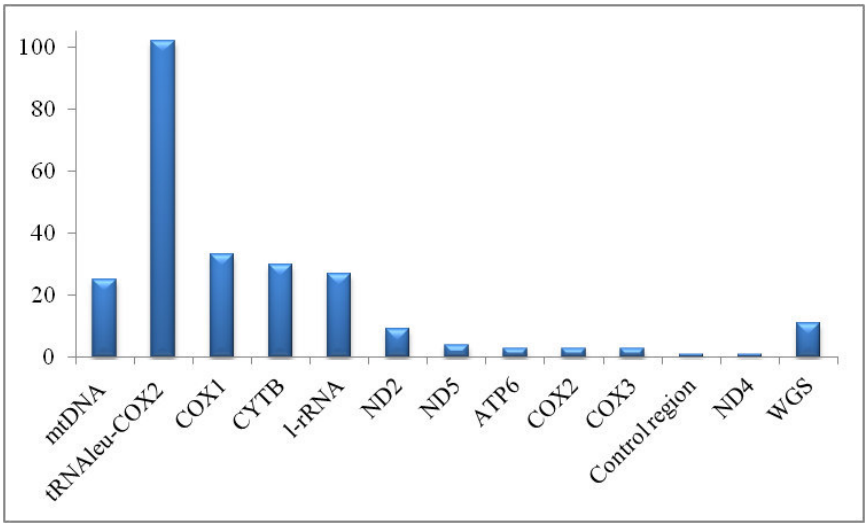


Figure II-6 – Percentage of publications using the entire molecule (mtDNA), different fragments of the mitogenome or recently the complete mitogenome (WGS). Publications registered in the Web of Science from 1981 to 2017 (Data retrieved 10.12.2017).

Recent advances in next generation sequencing has made it possible to quickly and economically generate whole mitochondrial genome (mitogenome) sequences, (Jacobsen *et al.*, 2012). The mitogenome has higher resolution than individual genes and has been applied to a variety of phylogenetic and phylogeographic studies to solve shallow evolutionary histories mainly in recent timescales (Feutry *et al.*, 2017; Feutry *et al.*, 2014; Gilbert *et al.*, 2008; Hong *et al.*, 2017; Keis *et al.*, 2013). In the honey bee, there are very few studies using the mitogenome, most of them (6 out of 11) focus on the description of the mitogenome of several subspecies (Eimanifar *et al.*, 2016; Eimanifar *et al.*, 2017a; Eimanifar *et al.*, 2017b). Other studies used the mitogenome to answer to specific questions. For instance, Mikheyev *et al.* (2015) studied the mitogenome to understand the evolution of honey bees when exposed to parasites. Wragg *et al.* (2017) used the mitogenome together with the nuclear genome to understand the adaptation in Reunion Island. Ilyasov *et al.* (2016) used 12 out of the 13 protein-coding genes to understand which one may differentiate honey bee subspecies belonging to the A, M, C and O lineages.

Unquestionably, mitochondrial markers have provided invaluable insights into the patterns and processes of honey bee diversity. However, it is important to keep in mind that the mitochondrial DNA represents only a small portion of organisms' genome (Galtier *et al.*, 2009) and

only the maternal component of genetic variation. It is therefore possible that it may not represent the true population history (Galtier *et al.*, 2009). To overcome this drawback, it is recommended to use both mtDNA and nuclear markers (Evans *et al.*, 2013; Meixner *et al.*, 2013).

Nuclear markers

Nuclear markers are important because they provide a more complete picture of the genetic variation of an organism. Currently, the most popular nuclear markers are single nucleotide polymorphism (SNPs) and microsatellites. In contrast to mtDNA, nuclear markers are biparentally inherited providing maternal and paternal genetic information (Evans *et al.*, 2013).

Microsatellites

Microsatellites are one of the most popular nuclear markers used in the study of genetic variation in honey bees. They consist on short motifs ranging from 1 to 6 bases, repeated between 4 to 100 or more times (Tautz, 1993). The term "microsatellite" was first introduced by Litt and Luty (1989), other alternative designations include variable number of tandem repeats (VNTR), short tandem repeats (STR) and simple sequence repeats (SSR) (Meixner *et al.*, 2013).

The reason behind the popularity of microsatellites lies in the fact that they are co-dominant (heterozygotes can be distinguished from homozygotes), abundant along the genome and so, they are easily amplified by PCR, and they are highly polymorphic. The high polymorphism is explained by mutation rates ranging between 10^{-3} and 10^{-6} per cell generation, which is up to 10 orders of magnitude greater than for SNPs (Gemayel *et al.*, 2012). This high mutation rate makes microsatellites suitable to infer population level events (Arif & Khan, 2009). The polymorphisms usually come from the addition or deletion of an entire repeat motif. Such repeated polymorphism can be formed by two different processes: strand-slippage replication and recombination. Strand-slippage replication is a DNA replication error by which mispairing occurs between the template and nascent strands. Recombination events, such as unequal crossing over and gene conversion, may also lead to contractions and expansions of microsatellites.

The first *A. mellifera* microsatellite *locus* was described in 1993 (Estoup *et al.*, 1993) and since then, this marker has been applied in numerous studies to infer the evolutionary history of populations (Cánovas *et al.*, 2011; Coroian *et al.*, 2014; Franck *et al.*, 1998; Miguel *et al.*, 2007; Pentek-Zakar *et al.*, 2015), as well as to address other biological aspects such as mating frequency (Palmer & Oldroyd, 2000). In addition, microsatellites have been the marker of choice used in

conservation centres to identify C-lineage ancestry in *A. m. mellifera* populations (Jensen *et al.*, 2005; Soland-Reckeweg *et al.*, 2009; Strange *et al.*, 2008). Despite the relevance of this marker, it has significant drawbacks. Microsatellites have a complex mutation pattern, which creates difficulties for populations-genetic analysis; for instance, due to the high mutation rate, two alleles can be identical by state but not by descendent, which in turn may obscure lineage differentiation. Technical problems such as PCR artifacts complicate the automated scoring of microsatellites (Schlötterer, 2004). Additionally, comparing results between laboratories requires cross-calibration because of the inconsistencies in allele size calling due to the different gel migration and fluorescent dyes used in automatic sequencing machines (Schlötterer, 2004).

SNPs

Single nucleotide polymorphism (SNP) is the most recent marker used to study the genetics of honey bees. As suggested by the acronym, SNPs consist on a single base change in a DNA sequence with the frequency of the least frequent allele of 1% or greater (Vignal *et al.*, 2002). Generally, there are four possible nucleotides at each position, but due to the low mutation rate, which is about 10^{-8} to 10^{-9} changes per nucleotide per generation, SNPs are usually bi-allelic. The bi-allelic nature of the SNPs can also be explained by the clear bias towards the transition substitutions (purine to purine or a pyrimidine to pyrimidine substitution), leading to the prevalence of two SNP types (Vignal *et al.*, 2002). One possible explanation to this substitution bias is the high spontaneous rate of deamination of 5-methyl cytosine to thymidine in the CpG dinucleotides (Vignal *et al.*, 2002). In addition, the effects of transversions are usually more severe since they change the chemical structure of the DNA (Maresso & Broeckel, 2008).

Initially, STRs were preferred over SNPs because of their multi-allelic nature and the high costs associated with SNP discovery and genotyping. Among the numerous approaches for SNP discovery, the simplest way was to perform direct sequencing of a DNA sequence, a procedure which at a large scale tended to be costly (Vignal *et al.*, 2002). Therefore, most of the large-scale studies of genetic diversity genotyped polymorphisms have been previously identified in other populations (Gilad *et al.*, 2009). These studies have low resolution and an uneven distribution of SNPs along the genome. Moreover, if the SNPs are identified in a small panel of individuals or if they are applied to a different population, there is a deviation of the expected allele frequency leading to an ascertainment bias.

SNP genotyping techniques

There are several methods for SNP genotyping, most of them involve sequence amplification to introduce specificity and increase the number of molecules for the allele discrimination step. Allele discrimination involves allele-specific biochemical reactions including primer extension (MassARRAY system of Agena BioScience™), hybridization (GeneChip®), oligonucleotide ligation/polymerase chain reaction (SNPlex™) or enzymatic cleavage (Invader®). Some methods use the combination of two or more discrimination approaches, such as the TaqMan® that combines hybridization and 5' nuclease activity of polymerase. Alternatively, BeadArray™ (Illumina, CA) combines hybridization with primer extension and ligation for generating an allele specific product. The final SNP genotyping step is allele detection which includes mass spectrometry (MassARRAY system of Agena BioScience™), fluorescence (TaqMan®, Invader®, BeadArray™) or chemiluminescence (Pyrosequencing™) (Chen & Sullivan, 2003; Kwok, 2001; Syvanen, 2001).

The choice of the genotyping technique depends on both the required number of SNPs and the sample size (Sobrino *et al.*, 2005). When a study needs few SNP markers to be genotyped in a very large sample size, methods that share the costs of the SNPs by many samples are especially useful. One technology that fits well in this scenario is the TaqMan nuclease assay. The mass spectrometry assay MassARRAY system of Agena BioScience™ is also an interesting technology for SNP genotyping. It measures the molecular weight of the products and has the capacity of multiplexing reactions through primer extension making it flexible, robust and well suited for high-throughput applications (Sobrino *et al.*, 2005). This technology is cost-effective in two different scenarios; when there is a small number of markers and a larger number of samples, or a larger number of markers and a small number of samples.

The SNPs in honey bees studies

With the continuous development of SNP genotyping technologies, SNPs are becoming very popular as an alternative nuclear marker to study populations' evolutionary events. SNPs are more abundant and widespread in the genome than microsatellites (Weinstock *et al.*, 2006), and they evolve in accordance to the infinite allele model which is more efficient than the stepwise mutation model (associated with STR) in the study of the origin of individuals. At a technical level, SNPs display lower genotyping errors, have higher quality data, are more amenable to automated

analysis and data interpretation and can be standardized, therefore allowing for experiments to be easily replicated between laboratories (reviewed by Vignal *et al.* (2002).

In honey bees, medium-density SNPs were mostly genotyped using Illumina technology with the aim of studying the origin of *Apis mellifera* L. (Whitfield *et al.*, 2006), the demographic history of the Iberian honey (Chávez-Galarza *et al.*, 2015), the search for signatures of selection between native and invasive *A. m. scutellata* populations (Zayed & Whitfield, 2008) as well as selection in Iberian populations (Chávez-Galarza *et al.*, 2013). In addition, SNPs are a powerful tool for testing the breeding stock in conservation centres and, unlike microsatellites, SNP-based genetic data can be readily incorporated in shared genetic databases, facilitating the implementation of conservation strategies. Pinto *et al.* (2014) used 1183 SNPs genotyped using Illumina's BeadArray Technology and the Illumina GoldenGate Assay with a custom Oligo Pool Assay to estimate C-lineage introgression in *A. m. mellifera* accross Europe. However, the costs associated with this SNP panel were still very high for this technique to be applied in conservation centres. To address this issue Muñoz *et al.* (2015) created a reduced panel containing the most ancestry-informative markers (AIMS) which was able to assign individuals to the correct origin and to calculate admixture levels with a high degree of accuracy.

With next generation sequencing (NGS) platforms, it is possible to identify thousands to millions of SNPs, in model and non-model organisms, directly from sequencing data (Baird *et al.*, 2008). This advance overcame some of the initial drawbacks of SNPs because now, instead of analysing limited genomic regions and few loci, the identified SNPs are distributed along entire genomes in functional and non-functional parts. Moreover, the ascertainment bias is no longer a problem since most variants, common and rare, can be discovered with the appropriate sequencing read coverage (Koboldt *et al.*, 2013).

Next generation sequencing

Next generation sequencing (NGS) platforms provide unprecedented capacity allowing sequencing the entire genome in a time and cost-effective manner. The development of a population-scale whole-genome-sequencing (WGS) dataset is greatly facilitated for the honey bee due to several reasons. Firstly, a reference genome is available. The *Apis mellifera* genome has been published in 2006 (Weinstock *et al.*, 2006) and has since then been frequently updated (Elsik *et al.*, 2014), thus making mapping or genome assemblage a lot easier. Secondly, the honey bee genome is only

236 Mbp long, which enables sufficient sequence coverage to obtain quality SNPs at an affordable cost. Thirdly, the honey bee haplodiploid system offers the possibility of sequencing haploid males thus prompting more accurate calls of polymorphic sites as well as preventing uncertainties of reconstructing haplotypes.

High-throughput sequencing and computational technologies coupled with increasingly sophisticated analytical tools have changed the scale of analysis from limited genomic regions and few loci to whole genomes. The millions of SNPs identified by WGS have been important to gain insights into honey bee population histories, as well as into the genetic mechanisms underlying adaptation and speciation. WGS have been used to shed light on the evolutionary history and origin of the honey bee. Wallberg *et al.* (2014) sampled 14 worldwide populations and did not find evidence for an African origin, while Cridland *et al.* (2017), using the Wallberg *et al.* (2014) and Harpur *et al.* (2014) genomes, proposed that the origin of *A. mellifera* lies in Africa, where substantial diversification occurred followed by radiations of the M lineage out of Africa into Europe, and of the C/O lineages back into the Middle East. While more populations need to be sampled to fully understand the origin of the honey bee, there is progress in understanding adaptive variation among honey bee populations, an important goal in the current context of a global human-mediated environmental crisis. The long-standing goal of uncovering the genetic basis of adaptation has never been so important because it will enable predictions on how organisms will respond to a rapidly changing world, which, in turn, will help design mitigating strategies. With WGS is possible to study almost all genetic variation present in a subset of individuals giving the opportunity to directly identify the genes and the mechanisms that allow the organism to adapt to different environments.

Different genomic regions potentially important for honey bee adaptation have been identified (Chen *et al.*, 2016; Fuller *et al.*, 2015; Harpur *et al.*, 2014; Nelson *et al.*, 2017; Wallberg *et al.*, 2014; Wallberg *et al.*, 2017; Wragg *et al.*, 2017). For instance, Wallberg *et al.* (2014) identified genes involved in sperm motility and immune system. Harpur *et al.* (2014) found that *Apoidea*- and *Apis*-specific genes are enriched for signatures of positive selection. Chen *et al.* (2016) compared honey bees from temperate and tropical regions and found candidate genes related to fat body and the Hippo signalling pathway. Fuller *et al.* (2015) sequenced 11 workers from Kenya and identified genes with signatures of positive selection that are involved in different biological processes such as metabolism, neuronal development, immunity, reproduction, gland

development and gland secretions cuticle formation. Parejo *et al.* (2017) identified signals of putative human-mediated selection arising due to different breeding practices in two managed populations . In addition, WGS have been important to identify the genes involved in the production of royal jelly (Wragg *et al.*, 2016), Varroa resistance (Haddad *et al.*, 2016), behavioural differences between scouts and recruits (Southey *et al.*, 2016), as well as adaptation to altitude (Wallberg *et al.*, 2017). In addition to the evolutionary history and adaptation, WGS data has further been used to understand the genomic architecture and evolution of the honey bee genome which is greatly influenced by the extremely high rates of meiotic recombination averaging 19–37 cM/Mb (Liu *et al.*, 2015; Wallberg *et al.*, 2015). Studying genetic recombination is important because it has a huge impact on the efficiency of natural selection on genetic variation and evolution in honey bees and appears to play a dominant role in genome evolution (Wallberg *et al.*, 2015).

Next generation sequencing tecnologies

Extraordinary progress has been made in genome sequencing technologies, which can be divided in first, second and third sequencing generations (Fuentes-Pardo & Ruzzante, 2017; Goodwin *et al.*, 2016). The first generation consists on the Sanger sequencing introduced in 1977 (Sanger *et al.*, 1977). This method is highly accurate (per-base accuracy of 99.999%), the length of the reads is around 1000 bp (Shendure & Ji, 2008). However, early on, it was clear that to answer to complex biological question on routine basis a different technology would be necessary. The Sanger method is a low-throughput technology and a total automation of sample preparation is very difficult (Ansorge, 2009). The second generation technologies or next generation sequencing (NGS) appeared between 2005 and 2010 and came to revolutionize the field of biology. These technologies increased in several orders of magnitude the speed of sequencing and decreased substantially the cost per base. This was possible thanks to some technical differences compared with Sanger technology, such as, instead of requiring bacterial cloning of DNA fragments, libraries are generated in a cell free system and the sequencing output is directly detected without the need for electrophoresis (Van Dijk *et al.*, 2014). However, different challenges arise: the huge amount of data generated by these systems (over a gigabase per run) requires enormous computing storage and more efficient computer algorithms (Ansorge, 2009). Another limitation is related with the short-read lengths (75-700 pb) which makes it difficult to assemble large genomes. However, different algorithms have been developed to manage short reads. The second generation

technologies require a PCR to increase the signal-to-noise ratio because the systems are not sensitive enough to detect the extension of one base at the individual DNA template molecule level (Buermans & Den Dunnen, 2014), but the PCR is a major source of bias due to its potential for distortion of reads abundance levels (Buermans & Den Dunnen, 2014). Finally the error rates are higher than in Sanger technology (accuracy >99.5%) (Fuentes-Pardo & Ruzzante, 2017). The first NGS technology to be released in 2005 was the pyrosequencing method by 454 Life Sciences (now Roche) that uses "sequencing-by-synthesis" system. One year later, the Solexa/Illumina sequencing platform was commercialized. The Oligo Ligation Detection (SOLiD) sequencing platform was released in 2007. Among them, the 454 was discontinued and Illumina is now the leading NGS platform offering the highest throughput and the lowest per-base cost. Other promising technologies, called the third generation, with the advantage of producing long reads (average ~2–10Kb) have recently been introduced. Examples of third generation technologies are Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (MinIONTM). These two are based on a "single-molecule real-time" method (Van Dijk *et al.*, 2014). Illuminas 10X Genomics based on synthetic long-reads sequencing is another method of this latest generation. The long reads are ideal to complete *de novo* assemblies, to reveal complex long range genomic structures, and to detect large structural variants. There are, however, several notable limitations. In "single-molecule real-time" methods, the error rate is high (15% -30%); and the PacBio and Illumina synthetic long-reads are more expensive than the standard Illumina platform; also, Illumina synthetic long-reads may exhibit lower accuracy when assembling highly heterozygous genomic regions (Kuleshov *et al.*, 2016).

References

Allendorf, F. W., & Luikart, G. (2007). Conservation and the genetics of populations. *Mammalia, 71*(4), 189-197. doi: doi.org/10.1515/MAMM.2007.038

Ansorge, W. J. (2009). Next-generation DNA sequencing techniques. *New Biotechnology, 25*(4), 195-203. doi: https://doi.org/10.1016/j.nbt.2008.12.009

Arias, M. C., Rinderer, T. E., & Sheppard, W. S. (2006). Further characterization of honey bees from the Iberian Peninsula by allozyme, morphometric and mtDNA haplotype analyses. *Journal of Apicultural Research, 45*(4), 188-196. doi: doi.org/10.1080/00218839.2006.11101346

Arias, M. C., & Sheppard, W. S. (2005). Phylogenetic relationships of honey bees (Hymenoptera: Apinae: Apini) inferred from nuclear and mitochondrial DNA sequence data. *Molecular Phylogenetics and Evolution, 37*(1), 25-35. doi: doi.org/10.1016/j.ympev.2005.02.017

Arif, I. A., & Khan, H. A. (2009). Molecular markers for biodiversity analysis of wildlife animals: a brief review. *Animal Biodiversity and Conservation, 32*(1), 9-17.

Avise, J. C., & Ellis, D. (1986). Mitochondrial DNA and the evolutionary genetics of higher animals [and Discussion]. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 312*(1154), 325-342. doi: 10.1098/rstb.1986.0011

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS One, 3*(10), e3376. doi: https://doi.org/10.1371/journal.pone.0003376

Ballard, J. W. O., & Whitlock, M. C. (2004). The incomplete natural history of mitochondria. *Molecular Ecology, 13*(4), 729-744. doi: 10.1046/j.1365-294X.2003.02063.x

Bertrand, B., Alburaki, M., Legout, H., Moulin, S., Mougel, F., & Garnery, L. (2015). MtDNA COI-COII marker and drone congregation area: An efficient method to establish and monitor honeybee (*Apis mellifera* L.) conservation centres. *Molecular Ecology Resources, 15*(3), 673-683. doi: 10.1111/1755-0998.12339

Bouga, M., Harizanis, P. C., Kilias, G., & Alahiotis, S. (2005). Genetic divergence and phylogenetic relationships of honey bee *Apis mellifera* (Hymenoptera: Apidae) populations from Greece and Cyprus using PCR – RFLP analysis of three mtDNA segments. *Apidologie, 36*(3), 335-344. doi: 10.1051/apido:2005021

Brodschneider, R., & Crailsheim, K. (2010). Nutrition and health in honey bees. *Apidologie, 41*(3), 278-294. doi: doi.org/10.1051/apido/2010012

Büchler, R., Costa, C., Hatjina, F., Andonov, S., Meixner, M. D., Le Conte, Y., Uzunov, A., Berg, S., Bienkowska, M., & Bouga, M. (2014). The influence of genetic origin and its interaction with environmental effects on the survival of *Apis mellifera* L. colonies in Europe. *Journal of Apicultural Research, 53*(2), 205-214. doi: 10.3896/IBRA.1.53.2.03

Buermans, H. P. J., & Den Dunnen, J. T. (2014). Next generation sequencing technology: advances and applications. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1842*(10), 1932-1941. doi: https://doi.org/10.1016/j.bbadis.2014.06.015

Cánovas, F., De la Rúa, P., Serrano, J., & Galián, J. (2008). Geographical patterns of mitochondrial DNA variation in *Apis mellifera iberiensis* (Hymenoptera: Apidae). *Journal of Zoological Systematics and Evolutionary Research, 46*(1), 24-30. doi: 10.1111/j.1439-0469.2007.00435.x

Cánovas, F., De la Rúa, P., Serrano, J., & Galián, J. (2011). Microsatellite variability reveals beekeeping influences on Iberian honeybee populations. *Apidologie, 42*(3), 235-251. doi: https://doi.org/10.1007/s13592-011-0020-1

Chávez-Galarza, J., Henriques, D., Johnston, J. S., Azevedo, J. C., Patton, J. C., Munoz, I., De la Rua, P., & Pinto, M. A. (2013). Signatures of selection in the Iberian honey bee (*Apis mellifera iberiensis*) revealed by a genome scan analysis of single nucleotide polymorphisms. *Molecular Ecology, 22*(23), 5890-5907. doi: 10.1111/mec.12537

Chávez-Galarza, J., Henriques, D., Johnston, J. S., Carneiro, M., Rufino, J., Patton, J. C., & Pinto, M. A. (2015). Revisiting the Iberian honey bee (*Apis mellifera iberiensis*) contact zone: maternal and genome-wide nuclear variations provide support for secondary contact from historical refugia. *Molecular Ecology, 24*(12), 2973-2992. doi: 10.1111/mec.13223

Chen, C., Liu, Z., Pan, Q., Chen, X., Wang, H., Guo, H., Liu, S., Lu, H., Tian, S., Li, R., & Shi, W. (2016). Genomic analyses reveal demographic history and temperate adaptation of the newly discovered honey bee subspecies *Apis mellifera sinisxinyuan n. ssp*. *Molecular Biology and Evolution, 33*(5), 1337-1348. doi: 10.1093/molbev/msw017

Chen, X., & Sullivan, P. F. (2003). Single nucleotide polymorphism genotyping: biochemistry, protocol, cost and throughput. *The Pharmacogenomics Journal, 3*(2), 77-96. doi: 10.1038/sj.tpj.6500167

Clarke, K. E., Rinderer, T. E., Franck, P., Quezada-Euan, J. G., & Oldroyd, B. P. (2002). The Africanization of honeybees (*Apis mellifera* L.) of the Yucatan: A study of a massive hybridization event across time. *Evolution, 56*(7), 1462-1474. doi: https://doi.org/10.1554/0014-3820(2002)056[1462:TAOHAM]2.0.CO;2

Clary, D. O., & Wolstenholme, D. R. (1985). The mitochondrial DNA molecule of *Drosophila yakuba*: nucleotide sequence, gene organization, and genetic code. *Journal of Molecular Evolution, 22*(3), 252-271. doi: https://doi.org/10.1007/BF02099755

Collet, T., Ferreira, K. M., Arias, M. C., Soares, A. E. E., & Del Lama, M. A. (2006). Genetic structure of Africanized honeybee populations (*Apis mellifera* L.) from Brazil and Uruguay viewed through mitochondrial DNA COI-COII patterns. *Heredity, 97*(5), 329-335. doi: 10.1038/sj.hdy.6800875

Cornuet, J., & Garnery, L. (1991). Mitochondrial DNA variability in honeybees and its phylogeographic implications. *Apidologie, 22*(6), 627-642. doi: 10.1051/apido:19910606

Coroian, C. O., Muñoz, I., Schlüns, E. A., Paniti-Teleky, O. R., Erler, S., Furdui, E. M., Mărghitaş, L. A., Dezmirean, D. S., Schlüns, H., De la Rúa, P., & Moritz, R. F. A. (2014). Climate rather than geography separates two European honeybee subspecies. *Molecular Ecology, 23*(9), 2353-2361. doi: 10.1111/mec.12731

Coulson, R. N., Pinto, M. A., Tchakerian, M. D., Baum, K. A., Rubink, W. L., & Johnston, J. S. (2005). Feral honey bees in pine forest landscapes of east Texas. *Forest Ecology and Management, 215*(1-3), 91-102. doi: 10.1016/j.foreco.2005.05.005

Crane, E. (1999). Recent research on the world history of beekeeping. *Bee World, 80*(4), 174-186. doi: doi.org/10.1080/0005772X.1999.11099453

Cridland, J. M., Tsutsui, N. D., & Ramírez, S. R. (2017). The complex demographic history and evolutionary origin of the western honey bee, *Apis mellifera*. *Genome Biology and Evolution, 9*(2), 457-472. doi: https://doi.org/10.1093/gbe/evx009

Crozier, R. H., & Crozier, Y. C. (1993). The mitochondrial genome of the honeybee *Apis mellifera*: complete sequence and genome organization. *Genetics, 133*(1), 97-117.

Crozier, Y. C., Koulianoss, S., & Crozier, R. H. (1991). An improved test for Africanized honeybee mitochondrial DNA. *Experientia, 47*(9), 968-969. doi: 10.1007/bf01929894

De la Rúa, P., Galian, J., Pedersen, B. V., & Serrano, J. (2006). Molecular characterization and population structure of *Apis mellifera* from Madeira and the Azores. *Apidologie, 37*(6), 699-708. doi: 10.1051/apido:2006044

De la Rúa, P., Jaffé, R., Dall'Olio, R., Muñoz, I., & Serrano, J. (2009). Biodiversity, conservation and current threats to European honeybees. *Apidologie, 40*(3), 263-284. doi: doi.org/10.1051/apido/2009027

De la Rúa, P., Jaffé, R., Muñoz, I., Serrano, J., Moritz, R. F. A., & Kraus, F. B. (2013). Conserving genetic diversity in the honeybee: Comments on Harpur et al.(2012). *Molecular Ecology, 22*(12), 3208-3210. doi: 10.1111/mec.12333

De la Rúa, P., Serrano, J., & Galian, J. (1998). Mitochondrial DNA variability in the Canary Islands honeybees (*Apis mellifera* L.). *Molecular Ecology, 7*(11), 1543-1547. doi: 10.1046/j.1365-294x.1998.00468.x

Dowling, T. E., & Secor, C. L. (1997). The role of hybridization and introgression in the diversification of animals. *Annual Review of Ecology and Systematics, 28*(1), 593-619. doi: doi.org/10.1146/annurev.ecolsys.28.1.593

Eimanifar, A., Kimball, R. T., Braun, E. L., & Ellis, J. D. (2016). The complete mitochondrial genome of the hybrid honey bee, *Apis mellifera capensis* x *Apis mellifera scutellata*, from South Africa. *Mitochondrial DNA Part B-Resources, 1*, 856-857. doi: 10.1080/23802359.2016.1250132

Eimanifar, A., Kimball, R. T., Braun, E. L., Fuchs, S., Grunewald, B., & Ellis, J. D. (2017a). The complete mitochondrial genome of *Apis mellifera meda* (Insecta: Hymenoptera: Apidae). *Mitochondrial DNA Part B-Resources, 2*, 268-269. doi: 10.1080/23802359.2017.1325342

Eimanifar, A., Kimball, R. T., Braun, E. L., Moustafa, D. M., Haddad, N., Fuchs, S., Grunewald, B., & Ellis, J. D. (2017b). The complete mitochondrial genome of the Egyptian honey bee, *Apis mellifera lamarckii* (Insecta: Hymenoptera: Apidae). *Mitochondrial DNA Part B-Resources, 2*, 270-272. doi: 10.1080/23802359.2017.1325343

Elsik, C. G., Worley, K. C., Bennett, A. K., Beye, M., Camara, F., Childers, C. P., de Graaf, D. C., Debyser, G., Deng, J., & Devreese, B. (2014). Finding the missing honey bee genes: lessons learned from a genome upgrade. *BioMed Central Genomics, 15*(1), 86. doi: https://doi.org/10.1186/1471-2164-15-86

Engel, M. S. (1999). The taxonomy of recent and fossil honey bees (Hymenoptera: Apidae; *Apis*). *Journal of Hymenoptera Research, 8* (2), 165-196.

Estoup, A., Garnery, L., Solignac, M., & Cornuet, J. (1995). Microsatellite variation in honey bee (*Apis mellifera* L.) populations: hierarchical genetic structure and test of the infinite allele and stepwise mutation models. *Genetics, 140*(2), 679-695.

Estoup, A., Solignac, M., Harry, M., & Cornuet, J. (1993). Characterization of (GT) n and (CT) n microsatellites in two insect species: *Apis mellifera* and *Bombus terrestris*. *Nucleic Acids Research, 21*(6), 1427-1431. doi: https://doi.org/10.1093/nar/21.6.1427

Evans, J. D., Schwarz, R. S., Chen, Y. P., Budge, G., Cornman, R. S., De la Rua, P., de Miranda, J. R., Foret, S., Foster, L., & Gauthier, L. (2013). Standard methods for molecular research in *Apis mellifera*. *Journal of Apicultural Research, 52*(4), 1-54. doi: doi.org/10.3896/IBRA.1.52.4.11

Feutry, P., Berry, O., Kyne, P. M., Pillans, R. D., Hillary, R. M., Grewe, P. M., Marthick, J. R., Johnson, G., Gunasekera, R. M., Bax, N. J., & Bravington, M. (2017). Inferring contemporary and historical genetic connectivity from juveniles. *Molecular Ecology, 26*(2), 444-456. doi: 10.1111/mec.13929

Feutry, P., Kyne, P. M., Pillans, R. D., Chen, X., Naylor, G. J. P., & Grewe, P. M. (2014). Mitogenomics of the Speartooth Shark challenges ten years of control region sequencing. *BioMed Central Evolutionary Biology, 14*(1), 232. doi: 10.1186/s12862-014-0232-x

Francis, R. M., Kryger, P., Meixner, M., Bouga, M., Ivanova, E., Andonov, S., Berg, S., Bienkowska, M., Büchler, R., & Charistos, L. (2014). The genetic origin of honey bee colonies used in the COLOSS genotype-environment interactions experiment: a comparison of methods. *Journal of Apicultural Research, 53*(2), 188-204. doi: 10.3896/IBRA.1.53.2.02

Franck, P., Garnery, L., Celebrano, G., Solignac, M., & Cornuet, J. (2000). Hybrid origins of honeybees from Italy (*Apis mellifera ligustica*) and Sicily (*A. m. sicula*). *Molecular Ecology, 9*(7), 907-921. doi: 10.1046/j.1365-294x.2000.00945.x

Franck, P., Garnery, L., Loiseau, A., Oldroyd, B. P., Hepburn, H. R., Solignac, M., & Cornuet, J. M. (2001). Genetic diversity of the honeybee in Africa: microsatellite and mitochondrial data. *Heredity, 86*(4), 420-430. doi: 10.1046/j.1365-2540.2001.00842.x

Franck, P., Garnery, L., Solignac, M., & Cornuet, J. M. (1998). The origin of west European subspecies of honeybees (*Apis mellifera*): New insights from microsatellite and mitochondrial data. *Evolution; International Journal of Organic Evolution., 52*(4), 1119-1134. doi: 10.2307/2411242

Fuentes-Pardo, A. P., & Ruzzante, D. E. (2017). Whole-genome sequencing approaches for conservation biology: advantages, limitations, and practical recommendations. *Molecular Ecology, 26*(20), 5369-5406. doi: 10.1111/mec.14264

Fuller, Z. L., Nino, E. L., Patch, H. M., Bedoya-Reina, O. C., Baumgarten, T., Muli, E., Mumoki, F., Ratan, A., McGraw, J., Frazier, M., Masiga, D., Schuster, S., Grozinger, C. M., & Miller, W. (2015). Genome-wide analysis of signatures of selection in populations of African honey bees (*Apis mellifera*) using new web-based tools. *BioMed Central Genomics, 16*(1), 518. doi: 10.1186/s12864-015-1712-0

Galtier, N., Nabholz, B., Glémin, S., & Hurst, G. D. D. (2009). Mitochondrial DNA as a marker of molecular diversity: a reappraisal. *Molecular Ecology, 18*(22), 4541-4550. doi: 10.1111/j.1365-294X.2009.04380.x

Garnery, L., Cornuet, J., & Solignac, M. (1992). Evolutionary history of the honey bee *Apis mellifera* inferred from mitochondrial DNA analysis. *Molecular Ecology, 1*(3), 145-154. doi: 10.1111/j.1365-294X.1992.tb00170.x

Garnery, L., Franck, P., Baudry, E., Vautrin, D., Cornuet, J., & Solignac, M. (1998). Genetic diversity of the west European honey bee (*Apis mellifera mellifera* and *A. m. iberica*). I. Mitochondrial DNA. *Genetics Selection Evolution, 30*, S31-S47. doi: 10.1051/gse:19980702

Garnery, L., Solignac, M., Celebrano, G., & Cornuet, J. (1993). A simple test using restricted PCR-amplified mitochondrial DNA to study the genetic structure of *Apis mellifera L. Experientia, 49*(11), 1016-1021. doi: 10.1007/bf02125651

Gemayel, R., Cho, J., Boeynaems, S., & Verstrepen, K. J. (2012). Beyond junk-variable tandem repeats as facilitators of rapid evolution of regulatory and coding sequences. *Genes, 3*(3), 461-480. doi: 10.3390/genes3030461

Gilad, Y., Pritchard, J. K., & Thornton, K. (2009). Characterizing natural variation using next-generation sequencing technologies. *Trends in Genetics, 25*(10), 463-471. doi: 10.1016/j.tig.2009.09.003

Gilbert, M. T. P., Drautz, D. I., Lesk, A. M., Ho, S. Y. W., Qi, J., Ratan, A., Hsu, C., Sher, A., Dalén, L., Götherström, A., Tomsho, L. P., Rendulic, S., Packard, M., Campos, P. F., Kuznetsova, T. V., Shidlovskiy, F., Tikhonov, A., Willerslev, E., Iacumin, P., Buigues, B., Ericson, P. G. P., Germonpré, M., Kosintsev, P., Nikolaev, V., Nowak-Kemp, M., Knight, J. R., Irzyk, G. P., Perbost, C. S.,

Fredrikson, K. M., Harkins, T. T., Sheridan, S., Miller, W., & Schuster, S. C. (2008). Intraspecific phylogenetic analysis of Siberian woolly mammoths using complete mitochondrial genomes. *Proceedings of the National Academy of Sciences of the United States of America, 105*(24), 8327-8332. doi: 10.1073/pnas.0802315105

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics, 17*(6), 333-351. doi: 10.1038/nrg.2016.49

Gruber, K., Schoning, C., Otte, M., Kinuthia, W., & Hasselmann, M. (2013). Distinct subspecies or phenotypic plasticity? Genetic and morphological differentiation of mountain honey bees in East Africa. *Ecology and Evolution, 3*(10), 3204-3218. doi: 10.1002/ece3.711

Guzmán-Novoa, E., Eccles, L., Calvete, Y., Mcgowan, J., Kelly, P. G., & Correa-Benítez, A. (2010). *Varroa destructor* is the main culprit for the death and reduced populations of overwintered honey bee (*Apis mellifera*) colonies in Ontario, Canada. *Apidologie, 41*(4), 443-450. doi: 10.1051/apido/2009076

Haddad, N., Mahmud Batainh, A., Suleiman Migdadi, O., Saini, D., Krishnamurthy, V., Parameswaran, S., & Alhamuri, Z. (2016). Next generation sequencing of *Apis mellifera syriaca* identifies genes for varroa resistance and beneficial bee keeping traits. *Insect Science, 23*(4), 579-590. doi: 10.1111/1744-7917.12205

Hahn, C., Bachmann, L., & Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Research, 41*(13), e129-e129. doi: 10.1093/nar/gkt371

Hall, H. G., & Muralidharan, K. (1989). Evidence from mitochondrial DNA that African honey bees spread as continuous maternal lineages. *Nature, 339*(6221), 211-213.

Hall, H. G., & Smith, D. R. (1991). Distinguishing African and European honeybee matrilines using amplified mitochondrial DNA. *Proceedings of the National Academy of Sciences, 88*(10), 4548-4552. doi: 10.1073/pnas.88.10.4548

Harpur, B. A., Kent, C. F., Molodtsova, D., Lebon, J. M., Alqarni, A. S., Owayss, A. A., & Zayed, A. (2014). Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. *Proceedings of the National Academy of Sciences U S A, 111*(7), 2614-2619. doi: 10.1073/pnas.1315506111

Hong, Y., Duo, H., Hong, J., Yang, J., Liu, S., Yu, L., & Yi, T. (2017). Resequencing and comparison of whole mitochondrial genome to gain insight into the evolutionary status of the Shennongjia golden snub-nosed monkey (SNJ *R. roxellana*). *Ecology and Evolution, 7*(12), 4456-4464. doi: 10.1002/ece3.3011

Ilyasov, R. A., Poskryakov, A. V., & Nikolenko, A. G. (2016). Seven genes of mitochondrial genome enabling differentiation of honeybee subspecies *Apis mellifera*. *Russian Journal of Genetics, 52*(10), 1062-1070. doi: 10.1134/s1022795416090064

Ivanova, E. N. (2010). Investigation on genetic variability in honey bee populations from Bulgaria, Greece and Serbia. *Biotechnology & Biotechnological Equipment, 24*(sup1), 385-389. doi: https://doi.org/10.1080/13102818.2010.10817869

Jacobsen, M. W., Hansen, M. M., Orlando, L., Bekkevold, D., Bernatchez, L., Willerslev, E., & Gilbert, M. T. P. (2012). Mitogenome sequencing reveals shallow evolutionary histories and recent divergence time between morphologically and ecologically distinct European whitefish (*Coregonus spp.*). *Molecular Ecology, 21*(11), 2727-2742. doi: 10.1111/j.1365-294X.2012.05561.x

Jensen, A. B., Palmer, K. A., Boomsma, J. J., & Pedersen, B. V. (2005). Varying degrees of *Apis mellifera ligustica* introgression in protected populations of the black honeybee, *Apis mellifera mellifera*, in northwest Europe. *Molecular Ecology, 14*(1), 93-106. doi: 10.1111/j.1365-294X.2004.02399.x

Jensen, A. B., & Pedersen, B. V. (2005b). Honeybee Conservation: a case story from Læsø island, Denmark *Beekeping and Conserving Biodiversity of Honeybees* (pp. 142-164): Northern Bee Books.

Johnson, R. M., Ellis, M. D., Mullin, C. A., & Frazier, M. (2010). Pesticides and honey bee toxicity–USA. *Apidologie, 41*(3), 312-331. doi: doi.org/10.1051/apido/2010018

Kasangaki, P., Nyamasyo, G., Ndegwa, P., Kajobe, R., Angiro, C., Kato, A., & Masembe, C. (2017). Mitochondrial DNA (mtDNA) markers reveal low genetic variation and the presence of two honey bee races in Uganda's agro-ecological zones. *Journal of Apicultural Research, 56*(2). doi: 10.1080/00218839.2017.1287997

Keis, M., Remm, J., Ho, S. Y. W., Davison, J., Tammeleht, E., Tumanov, I. L., Saveljev, A. P., Männil, P., Kojola, I., Abramov, A. V., Margus, T., & Saarma, U. (2013). Complete mitochondrial genomes and a novel spatial genetic method reveal cryptic phylogeographical structure and migration patterns among brown bears in north-western Eurasia. *Journal of Biogeography, 40*(5), 915-927. doi: 10.1111/jbi.12043

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., & Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell, 155*(1), 27-38. doi: https://doi.org/10.1016/j.cell.2013.09.006

Kotthoff, U., Wappler, T., & Engel, M. S. (2013). Greater past disparity and diversity hints at ancient migrations of European honey bee lineages into Africa and Asia. *Journal of Biogeography, 40*(10), 1832-1838. doi: 10.1111/jbi.12151

Kuleshov, V., Snyder, M. P., & Batzoglou, S. (2016). Genome assembly from synthetic long read clouds. *Bioinformatics, 32*(12), i216-i224. doi: https://doi.org/10.1093/bioinformatics/btw267

Kwok, P. (2001). Methods for genotyping single nucleotide polymorphisms. *Annual review of genomics and human genetics, 2*(1), 235-258. doi: https://doi.org/10.1146/annurev.genom.2.1.235

Lansman, R. A., Shade, R. O., Shapira, J. F., & Avise, J. C. (1981). The use of restriction endonucleases to measure mitochondrial DNA sequence relatedness in natural populations. *Journal of Molecular Evolution, 17*(4), 214-226. doi: doi.org/10.1007/BF01732759

Le Conte, Y., & Navajas, M. (2008). Climate change: impact on honey bee populations and diseases. *Revue Scientifique et Technique-Office International des Epizooties, 27*(2), 499-510.

Litt, M., & Luty, J. A. (1989). A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *American journal of human genetics, 44*(3), 397.

Liu, H., Zhang, X., Huang, J., Chen, J., Tian, D., Hurst, L. D., & Yang, S. (2015). Causes and consequences of crossing-over evidenced via a high-resolution recombinational landscape of the honey bee. *Genome biology, 16*(1), 15. doi: https://doi.org/10.1186/s13059-014-0566-0

Lynch, M., & Walsh, B. (2007). *The origins of genome architecture* (Vol. 98): Sinauer Associates Sunderland (MA).

Maresso, K., & Broeckel, U. (2008). Genotyping Platforms for Mass-Throughput Genotyping with SNPs, Including Human Genome-Wide Scans. *Advances in genetics, 60*, 107-139 %@ 0065-2660. doi: https://doi.org/10.1016/S0065-2660(07)00405-1

McMahon, D. P., Natsopoulou, M. E., Doublet, V., Fürst, M., Weging, S., Brown, M. J., Gogol-Döring, A., & Paxton, R. J. (2016). *Elevated virulence of an emerging viral genotype as a driver of honeybee loss.* Paper presented at the Proceeding of the Royal Society B.

Meixner, M. D. (2010). A historical review of managed honey bee populations in Europe and the United States and the factors that may affect them. *Journal of invertebrate pathology, 103*, S80-S95. doi: doi.org/10.1016/j.jip.2009.06.011

Meixner, M. D., Leta, M. A., Koeniger, N., & Fuchs, S. (2011). The honey bees of Ethiopia represent a new subspecies of *Apis mellifera- Apis mellifera simensis* n. ssp. . *Apidologie, 42*, 425-437. doi: 10.1007/s13592-011-0007-y

Meixner, M. D., Pinto, M. A., Bouga, M., Kryger, P., Ivanova, E., & Fuchs, S. (2013). Standard methods for characterising subspecies and ecotypes of *Apis mellifera*. *Journal of Apicultural Research, 52*(4), 1-28. doi: doi.org/10.3896/IBRA.1.52.4.05

Miguel, I., Iriondo, M., Garnery, L., Sheppard, W. S., & Estonba, A. (2007). Gene flow within the M evolutionary lineage of *Apis mellifera*: role of the Pyrenees, isolation by distance and post-glacial re-colonization routes in the western Europe. *Apidologie, 38*(2), 141-155. doi: 10.1051/apido:2007007

Mikheyev, A. S., Tin, M. M. Y., Arora, J., & Seeley, T. D. (2015). Museum samples reveal rapid evolution by wild honey bees exposed to a novel parasite. *Nature Communications, 6*. doi: 10.1038/ncomms8991

Mortensen, A. N., & Ellis, J. D. (2015). The frequency of African (*Apis mellifera scutellata* Lepeletier) matrilineal usurpation of managed European-derived honey bee (*A. mellifera* L.) colonies in the southeastern United States. *Insectes Sociaux, 62*(2), 151-155. doi: 10.1007/s00040-014-0383-1

Muñoz, I., Henriques, D., Jara, L., Johnston, J. S., Chávez-Galarza, J., De La Rúa, P., & Pinto, M. A. (2017). SNPs selected by information content outperform randomly selected microsatellite loci for delineating genetic identification and introgression in the endangered Dark European honeybee (*Apis mellifera mellifera*). *Mol Ecol Resour, 17*(4), 783-795. doi: 10.1111/1755-0998.12637

Muñoz, I., Henriques, D., Johnston, J. S., Chavez-Galarza, J., Kryger, P., & Pinto, M. A. (2015). Reduced SNP panels for genetic identification and introgression analysis in the Dark honey bee (*Apis mellifera mellifera*). *PLoS One, 10*(4), e0124365. doi: 10.1371/journal.pone.0124365

Mutinelli, F., Montarsi, F., Federico, G., Granato, A., Ponti, A. M., Grandinetti, G., Ferrè, N., Franco, S., Duquesne, V., & Rivière, M. (2014). Detection of *Aethina tumida* Murray (Coleoptera: Nitidulidae.) in Italy: outbreaks and early reaction measures. *Journal of Apicultural Research, 53*(5), 569-575. doi: doi.org/10.3896/IBRA.1.53.5.13

Nelson, R. M., Wallberg, A., Simões, Z. L. P., Lawson, D. J., & Webster, M. T. (2017). Genome-wide analysis of admixture and adaptation in the Africanized honeybee. *Molecular Ecology, 26*, 3603–3617. doi: 10.1111/mec.14122

Neumann, P., & Blacquière, T. (2017). The Darwin cure for apiculture? Natural selection and managed honeybee health. *Evolutionary applications, 10*(3), 226-230. doi: 10.1111/eva.12448

Nielsen, D. I., Ebert, P. R., Page, R. E., Hunt, G. J., & Guzmán-Novoa, E. (2000). Improved polymerase chain reaction-based mitochondrial genotype assay for identification of the Africanized honey bee (Hymenoptera: Apidae). *Annals of the Entomological Society of America, 93*(1), 1-6. doi: https://doi.org/10.1603/0013-8746(2000)093[0001:IPCRBM]2.0.CO;2

Nolte, A. W., & Tautz, D. (2010). Understanding the onset of hybrid speciation. *Trends in Genetics, 26*(2), 54-58. doi: doi.org/10.1016/j.tig.2009.12.001

Palmer, K. A., & Oldroyd, B. P. (2000). Evolution of multiple mating in the genus *Apis*. *Apidologie, 31*(2), 235-248. doi: 10.1051/apido:2000119

Parejo, M., Henriques, D., Pinto, M. A., Soland-Reckewege, G., & Neuditschkoa, M. (2018). Empirical comparison of microsatellite and SNP markers to estimate introgression in *Apis mellifera mellifera*. *Journal of Apicultural Research, submitted*.

Parejo, M., Wragg, D., Gauthier, L., Vignal, A., Neumann, P., & Neuditschko, M. (2016). Using Whole-Genome Sequence Information to Foster Conservation Efforts for the European Dark Honey Bee, *Apis mellifera mellifera*. *Frontiers in Ecology and Evolution, 4*, 140 doi: 10.3389/fevo.2016.00140

Parejo, M., Wragg, D., Henriques, D., Vignal, A., & Neuditschko, M. (2017). Genome-wide scans between two honeybee populations reveal putative signatures of human-mediated selection. *Animal genetics, 48*(6), 704-707. doi: 10.1111/age.12599

Pentek-Zakar, E., Oleksa, A., Borowik, T., & Kusza, S. (2015). Population structure of honey bees in the Carpathian Basin (Hungary) confirms introgression from surrounding subspecies. *Ecology and Evolution, 5*(23), 5456-5467. doi: 10.1002/ece3.1781

Pinto, M. A., Henriques, D., Chávez-Galarza, J., Kryger, P., Garnery, L., van der Zee, R., Dahle, B., Soland-Reckeweg, G., De la Rúa, P., Dall' Olio, R., Carreck, N. L., & Johnston, J. S. (2014). Genetic integrity of the Dark European honey bee (*Apis mellifera mellifera*) from protected populations: a genome-wide assessment using SNPs and mtDNA sequence data. *Journal of Apicultural Research, 53*(2), 269-278. doi: 10.3896/ibra.1.53.2.08

Pinto, M. A., Henriques, D., Neto, M., Guedes, H., Munoz, I., Azevedo, J. C., & De la Rúa, P. (2013). Maternal diversity patterns of Ibero-Atlantic populations reveal further complexity of Iberian honeybees. *Apidologie, 44*(4), 430-439. doi: 10.1007/s13592-013-0192-y

Pinto, M. A., Johnston, J. S., Rubink, W. L., Coulson, R. N., Patton, J. C., & Sheppard, W. S. (2003). Identification of Africanized honey bee (Hymenoptera : Apidae) mitochondrial DNA: Validation of a rapid polymerase chain reaction-based assay. *Annals of the Entomological Society of America, 96*(5), 679-684. doi: 10.1603/0013-8746(2003)096[0679:ioahbh]2.0.co;2

Pinto, M. A., Munoz, I., Chavez-Galarza, J., & De la Rúa, P. (2012). The Atlantic side of the Iberian Peninsula: a hot-spot of novel African honey bee maternal diversity. *Apidologie, 43*(6), 663-673. doi: 10.1007/s13592-012-0141-1

Pinto, M. A., Rubink, W. L., Coulson, R. N., Patton, J. C., & Johnston, J. S. (2004). Temporal pattern of Africanization in a feral honeybee population from Texas inferred from mitochondrial DNA. *Evolution, 58*(5), 1047-1055. doi: https://doi.org/10.1554/03-334

Pinto, M. A., Sheppard, W. S., Johnston, J. S., Rubink, W. L., Coulson, R. N., Schiff, N. M., Kandemir, I., & Patton, J. C. (2007). Honey bees (Hymenoptera : Apidae) of African origin exist in non-Africanized areas of the Southern United States: Evidence from mitochondrial DNA. *Annals of the Entomological Society of America, 100*(2), 289-295. doi: 10.1603/0013-8746(2007)100[289:hbhaoa]2.0.co;2

Prada, C. F., Duran, J. T., Salamanca, G., & Del Lama, M. A. (2009). Population genetics of *Apis mellifera* L. (Hymenoptera: Apidae) from Colombia. *Journal of Apicultural Research, 48*(1), 3-10. doi: 10.3896/ibra.1.48.1.02

Raffiudin, R., & Crozier, R. H. (2007). Phylogenetic analysis of honey bee behavioral evolution. *Molecular phylogenetics and evolution, 43*(2), 543-552. doi: doi.org/10.1016/j.ympev.2006.10.013

Rangel, J., Giresi, M., Pinto, M. A., Baum, K. A., Rubink, W. L., Coulson, R. N., & Johnston, J. S. (2016). Africanization of a feral honey bee (*Apis mellifera*) population in South Texas: does a decade make a difference? *Ecology and Evolution, 6*(7), 2158-2169. doi: 10.1002/ece3.1974

Rhymer, J. M., & Simberloff, D. (1996). Extinction by hybridization and introgression. *Annual Review of Ecology and Systematics, 27*(1), 83-109 doi: doi.org/10.1146/annurev.ecolsys.27.1.83

Rosenkranz, P., Aumeier, P., & Ziegelmann, B. (2010). Biology and control of *Varroa destructor*. *Journal of invertebrate pathology, 103*, S96-S119. doi: doi.org/10.1016/j.jip.2009.07.016

Ruttner, F. (1988). *Biogeography and taxonomy of honeybees*: Springer Science & Business Media.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences, 74*(12), 5463-5467.

Schäfer, M. O., Ritter, W., Pettis, J. S., & Neumann, P. r. (2010). Winter losses of honeybee colonies (Hymenoptera: Apidae): The role of infestations with *Aethina tumida* (Coleoptera: Nitidulidae) and *Varroa destructor* (Parasitiformes: Varroidae). *Journal of economic entomology, 103*(1), 10-16. doi: doi.org/10.1603/EC09233

Schiff, N. M., & Sheppard, W. S. (1993). Mitochondrial DNA evidence for the 19th century introduction of African honey bees into the United States. *Cellular and Molecular Life Sciences, 49*(6), 530-532. doi: https://doi.org/10.1007/BF01955156

Schiff, N. M., & Sheppard, W. S. (1996). Genetic differentiation in the queen breeding population of the western United States. *Apidologie, 27*(2), 77-86. doi: 10.1051/apido:19960202

Schlötterer, C. (2004). The evolution of molecular markers—just a matter of fashion? *Nature reviews genetics, 5*(1), 63-69. doi: doi:10.1038/nrg1249

Shaibi, T., Munoz, I., Dall'Olio, R., Lodesani, M., De la Rua, P., & Moritz, R. F. A. (2009). *Apis mellifera* evolutionary lineages in Northern Africa: Libya, where orient meets occident. *Insectes Sociaux, 56*(3), 293-300. doi: 10.1007/s00040-009-0023-3

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology, 26*(10), 1135-1145. doi: 10.1038/nbt1486

Sheppard, W. S., Rinderer, T. E., Garnery, L., & Shimanuki, H. (1999). Analysis of Africanized honey bee mitochondrial DNA reveals further diversity of origin. *Genetics and Molecular Biology, 22*(1), 73-75. doi: 10.1590/s1415-47571999000100015

Sheppard, W. S., Soares, A. E. E., DeJong, D., & Shimanuki, H. (1991). Hybrid status of honey bee populations near the historic origin of Africanization in Brazil. *Apidologie, 22*(6), 643-652. doi: doi.org/10.1051/apido:19910607

Sinacori, A., Rinderer, T. E., Lancaster, V., & Sheppard, W. S. (1998). A morphological and mitochondrial assessment of *Apis mellifera* from Palermo, Italy. *Apidologie, 29*(6), 481-490. doi: 10.1051/apido:19980601

Smith, D. R., & Brown, W. M. (1988). Polymorphisms in mitochondrial DNA of European and Africanized honeybees (*Apis mellifera*). *Cellular and Molecular Life Sciences, 44*(3), 257-260. doi: doi.org/10.1007/BF01941730

Smith, D. R., & Glenn, T. C. (1995). Allozyme polymorphisms in Spanish honeybees (*Apis mellifera iberica*). *The Journal of Heredity, 86*(1), 12-16. doi: doi.org/10.1093/oxfordjournals.jhered.a111518

Smith, D. R., Palopoli, M. F., Taylor, B. R., Garnery, L., Cornuet, J. M., Solignac, M., & Brown, W. M. (1991). Geographical overlap of two mitochondrial genomes in Spanish honeybees (*Apis mellifera iberica*). *The Journal of Heredity, 82*(2), 96-100. doi: doi.org/10.1093/oxfordjournals.jhered.a111062

Sobrino, B., Brión, M., & Carracedo, A. (2005). SNPs in forensic genetics: a review on SNP typing methodologies. *Forensic Science International, 154*(2), 181-194. doi: http://dx.doi.org/10.1016/j.forsciint.2004.10.020

Soland-Reckeweg, G., Heckel, G., Neumann, P., Fluri, P., & Excoffier, L. (2009). Gene flow in admixed populations and implications for the conservation of the Western honeybee, *Apis mellifera*. *Journal of Insect Conservation, 13*(3), 317. doi: doi.org/10.1007/s10841-008-9175-0

Southey, B. R., Zhu, P., Carr-Markell, M. K., Liang, Z. S., Zayed, A., Li, R., Robinson, G. E., & Rodriguez-Zas, S. L. (2016). Characterization of genomic variants associated with scout and recruit behavioral castes in honey bees using whole-genome sequencing. *PloS one, 11*(1), e0146430. doi: https://doi.org/10.1371/journal.pone.0146430

Stevanovic, J., Stanimirovic, Z., Radakovic, M., & Kovacevic, S. R. (2010). Biogeographic study of the honey bee (*Apis mellifera* L.) from Serbia, Bosnia and Herzegovina and Republic of Macedonia based on mitochondrial DNA analyses. *Russian Journal of Genetics, 46*(5), 603-609. doi: https://doi.org/10.1134/S1022795410050145

Strange, J. P., Garnery, L., & Sheppard, W. S. (2008). Morphological and molecular characterization of the Landes honey bee (*Apis mellifera* L.) ecotype for genetic conservation. *Journal of Insect Conservation, 12*(5), 527-537. doi: https://doi.org/10.1007/s10841-007-9093-6

Syvanen, A. (2001). Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Reviews Genetics, 2*(12), 930-942. doi: 10.1038/35103535

Tautz, D. (1993). Notes on the definition and nomenclature of tandemly repetitive DNA sequences *DNA fingerprinting: State of the science* (pp. 21-28): Springer.

Techer, M. A., Clemencet, J., Simiand, C., Preeaduth, S., Azali, H. A., Reynaud, B., & Delatte, H. (2017). Large-scale mitochondrial DNA analysis of native honey bee *Apis mellifera* populations reveals a new African subgroup private to the South West Indian Ocean islands. *BMC Genetics, 18*. doi: 10.1186/s12863-017-0520-8

Techer, M. A., Clemencet, J., Turpin, P., Volbert, N., Reynaud, B., & Delatte, H. (2015). Genetic characterization of the honeybee (*Apis mellifera*) population of Rodrigues Island, based on microsatellite and mitochondrial DNA. *Apidologie, 46*(4), 445-454. doi: 10.1007/s13592-014-0335-9

Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in genetics, 30*(9), 418-426. doi: https://doi.org/10.1016/j.tig.2014.07.001

vanEngelsdorp, D., & Meixner, M. D. (2010). A historical review of managed honey bee populations in Europe and the United States and the factors that may affect them. *Journal of Invertebrate Pathology, 103, Supplement*, S80-S95. doi: 10.1016/j.jip.2009.06.011

Vignal, A., Milan, D., SanCristobal, M., & Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution, 34*(3), 275. doi: https://doi.org/10.1186/1297-9686-34-3-275

Wallberg, A., Glemin, S., & Webster, M. T. (2015). Extreme recombination frequencies shape genome variation and evolution in the honeybee, *Apis mellifera*. *PLoS Genet, 11*(4), e1005189. doi: 10.1371/journal.pgen.1005189

Wallberg, A., Han, F., Wellhagen, G., Dahle, B., Kawata, M., Haddad, N., Simoes, Z. L., Allsopp, M. H., Kandemir, I., De la Rua, P., Pirk, C. W., & Webster, M. T. (2014). A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee *Apis mellifera*. *Nature Genetics, 46*(10), 1081-1088. doi: 10.1038/ng.3077

Wallberg, A., Schöning, C., Webster, M. T., & Hasselmann, M. (2017). Two extended haplotype blocks are associated with adaptation to high altitude habitats in East African honey bees. *PLoS Genetics, 13*(5), e1006792. doi: https://doi.org/10.1371/journal.pgen.1006792

Weinstock, G. M., Robinson, G. E., Gibbs, R. A., Worley, K. C., Evans, J. D., Maleszka, R., Robertson, H. M., Weaver, D. B., Beye, M., Bork, P., Elsik, C. G., Hartfelder, K., Hunt, G. J., Zdobnov, E. M., Amdam, G. V., Bitondi, M. M. G., Collins, A. M., Cristino, A. S., Lattorff, H. M. G., Lobo, C. H., Moritz, R. F. A., Nunes, F. M. F., Page Jr, R. E., Simões, Z. L. P., Wheeler, D., Carninci, P., Fukuda, S., Hayashizaki, Y., Kai, C., Kawai, J., Sakazume, N., Sasaki, D., Tagami, M., Albert, S., Baggerman, G., Beggs, K. T., Bloch, G., Cazzamali, G., Cohen, M., Drapeau, M. D., Eisenhardt, D., Emore, C.,

Ewing, M. A., Fahrbach, S. E., Forêt, S., Grimmelikhuijzen, C. J. P., Hauser, F., Hummon, A. B., Huybrechts, J., Jones, A. K., Kadowaki, T., Kaplan, N., Kucharski, R., Leboulle, G., Linial, M., Littleton, J. T., Mercer, A. R., Richmond, T. A., Rodriguez-Zas, S. L., Rubin, E. B., Sattelle, D. B., Schlipalius, D., Schoofs, L., Shemesh, Y., Sweedler, J. V., Velarde, R., Verleyen, P., Vierstraete, E., Williamson, M. R., Ament, S. A., Brown, S. J., Corona, M., Dearden, P. K., Dunn, W. A., Elekonich, M. M., Fujiyuki, T., Gattermeier, I., Gempe, T., Hasselmann, M., Kadowaki, T., Kage, E., Kamikouchi, A., Kubo, T., Kucharski, R., Kunieda, T., Lorenzen, M., Milshina, N. V., Morioka, M., Ohashi, K., Overbeek, R., Ross, C. A., Schioett, M., Shippy, T., Takeuchi, H., Toth, A. L., Willis, J. H., Wilson, M. J., Gordon, K. H. J., Letunic, I., Hackett, K., Peterson, J., Felsenfeld, A., Guyer, M., Solignac, M., Agarwala, R., Cornuet, J. M., Monnerot, M., Mougel, F., Reese, J. T., Schlipalius, D., Vautrin, D., Gillespie, J. J., Cannone, J. J., Gutell, R. R., Johnston, J. S., Eisen, M. B., Iyer, V. N., Iyer, V., Kosarev, P., Mackey, A. J., Solovyev, V., Souvorov, A., Aronstein, K. A., Bilikova, K., Chen, Y. P., Clark, A. G., Decanini, L. I., Gelbart, W. M., Hetru, C., Hultmark, D., Imler, J. L., Jiang, H., Kanost, M., Kimura, K., Lazzaro, B. P., Lopez, D. L., Simuth, J., Thompson, G. J., Zou, Z., De Jong, P., Sodergren, E., Csurös, M., Milosavljevic, A., Osoegawa, K., Richards, S., Shu, C. L., Duret, L., Elhaik, E., Graur, D., Anzola, J. M., Campbell, K. S., Childs, K. L., Collinge, D., Crosby, M. A., Dickens, C. M., Grametes, L. S., Grozinger, C. M., Jones, P. L., Jorda, M., Ling, X., Matthews, B. B., Miller, J., Mizzen, C., Peinado, M. A., Reid, J. G., Russo, S. M., Schroeder, A. J., St. Pierre, S. E., Wang, Y., Zhou, P., Jiang, H., Kitts, P., Ruef, B., Venkatraman, A., Zhang, L., Aquino-Perez, G., Whitfield, C. W., Behura, S. K., Berlocher, S. H., Sheppard, W. S., Smith, D. R., Suarez, A. V., Tsutsui, N. D., Wei, X., Wheeler, D., Havlak, P., Li, B., Liu, Y., Jovilet, A., Lee, S., Nazareth, L. V., Pu, L. L., Thorn, R., Stolc, V., Newman, T., Samanta, M., Tongprasit, W. A., Claudianos, C., Berenbaum, M. R., Biswas, S., De Graaf, D. C., Feyereisen, R., Johnson, R. M., Oakeshott, J. G., Ranson, H., Schuler, M. A., Muzny, D., Chacko, J., Davis, C., Dinh, H., Gill, R., Hernandez, J., Hines, S., Hume, J., Jackson, L., Kovar, C., Lewis, L., Miner, G., Morgan, M., Nguyen, N., Okwuonu, G., Paul, H., Santibanez, J., Savery, G., Svatek, A., Villasana, D., & Wright, R. (2006). Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature, 443*(7114), 931-949. doi: 10.1038/nature05260

Whitfield, C. W., Behura, S. K., Berlocher, S. H., Clark, A. G., Johnston, J. S., Sheppard, W. S., Smith, D. R., Suarez, A. V., Weaver, D., & Tsutsui, N. D. (2006). Thrice out of Africa: ancient and recent expansions of the honey bee, *Apis mellifera*. *Science, 314*(5799), 642-645. doi: 10.1126/science.1132772

Wragg, D., Marti-Marimon, M., Basso, B., Bidanel, J., Labarthe, E., Bouchez, O., Le Conte, Y., & Vignal, A. (2016). Whole-genome resequencing of honeybee drones to detect genomic selection in a population managed for royal jelly. *Scientific reports, 6*. doi: 10.1038/srep27168

Wragg, D., Techer, M. A., Canale-Tabet, K., Basso, B., Bidanel, J., Labarthe, E., Bouchez, O., Le Conte, Y., Clémencet, J., & Delatte, H. (2017). Autosomal and mitochondrial adaptation following admixture: A case study on the honeybees of Reunion Island. *Genome Biology and Evolution, 10*(1), 220-238. doi: https://doi.org/10.1093/gbe/evx247

Zayed, A., & Whitfield, C. W. (2008). A genome-wide signature of positive selection in ancient and recent invasive expansions of the honey bee *Apis mellifera*. *Proceedings of the National Academy of Sciences, 105*(9), 3421-3426. doi: 10.1073/pnas.0800107105

Chapter III.

Reduced SNP Panels for Genetic Identification and Introgression Analysis in the Dark Honey Bee (*Apis mellifera mellifera*)

The work presented in this chapter has been published:

Muñoz, I., **Henriques, D.**, Johnston, J. S., Chávez-Galarza, J., Kryger, P., & Pinto, M. A. (2015). Reduced SNP panels for genetic identification and introgression analysis in the Dark honey bee (*Apis mellifera mellifera*). *PloS one*, *10*(4), e0124365.

Abstract

Beekeeping activities, especially queen trading, have shaped the distribution of honey bee (*Apis mellifera*) subspecies in Europe, and have resulted in extensive introductions of two eastern European C-lineage subspecies (*A. m. ligustica* and *A. m. carnica*) into the native range of the M-lineage *A. m. mellifera* subspecies in Western Europe. As a consequence, replacement and gene flow between native and commercial populations have occurred at varying levels across western European populations. Genetic identification and introgression analysis using molecular markers is an important tool for management and conservation of honey bee subspecies. Previous studies have monitored introgression by using microsatellite, PCR-RFLP markers and most recently, high density assays using single nucleotide polymorphism (SNP) markers. While the latter are almost prohibitively expensive, the information gained to date can be exploited to create a reduced panel containing the most ancestry-informative markers (AIMs) for those purposes with very little loss of information. The objective of this study was to design reduced panels of AIMs to verify the origin of *A. m. mellifera* individuals and to provide accurate estimates of the level of C-lineage introgression into their genome. The discriminant power of the SNPs using a variety of metrics and approaches including the Weir & Cockerham's F_{ST}, an F_{ST}-based outlier test, Delta, informativeness (I_n), and PCA was evaluated. This study shows that reduced AIMs panels assign individuals to the correct origin and calculates the admixture level with a high degree of accuracy. These panels provide an essential tool in Europe for genetic stock identification and estimation of admixture levels which can assist management strategies and monitor honey bee conservation programs.

Introduction

The role of introgression and admixture in conservation is a dilemma: While natural admixture may be an important evolutionary force in speciation and maintenance of genetic diversity (Dowling & Secor, 1997; Nolte & Tautz, 2010), admixture induced by human activities may contribute, either directly or indirectly, to the extinction of many taxa (Rhymer & Simberloff, 1996). Introduction of species, subspecies and habitat modifications has caused increased rates of admixture with native flora and fauna and introgression that can generate extinction and irretrievable loss of combinations of genotypes throughout the entire genome (Allendorf & Luikart, 2007).

The honey bee, *Apis mellifera* L., represents a valuable model to study human-mediated change. Beekeeping has been practiced in Europe for many centuries (Crane, 2013), which has led to loss of native genetic diversity through three major mechanisms: (i) replacement of native populations by human-selected more docile and productive colonies, (ii) spread of honey bee pests and parasites, such as the mite *Varroa destructor* and the microsporidian *Nosema ceranae,* that have contributed to worldwide population declines (Nolte & Tautz, 2010; vanEngelsdorp & Meixner, 2010), and (iii) recurrent introductions of commercial colonies (reviewed by De la Rúa *et al.*, 2009).

The genetic diversity harbored in native honey bee subspecies is amongst the most important legacies that we can leave to future generations of beekeepers and farmers (Meixner *et al.*, 2010; Pinto *et al.*, 2014). Native honey bee subspecies are important reservoirs of local adaptations; their extinction means the loss of unique combinations of traits shaped by natural selection over extended periods of time. These combinations can be important for a more sustainable beekeeping, as shown by a recent pan-European experiment (Büchler *et al.*, 2014).

In Europe, honey bees show considerable differences in morphological, behavioural and biological characters across their range as a result of historical patterns of isolation and adaptation to environmental conditions (De la Rúa *et al.*, 2009). Those differences are materialized in 10 extant European subspecies, among the 30 subspecies currently recognized worldwide (Engel, 1999; Hepburn & Radloff, 1988; Meixner *et al.*, 2011; Ruttner, 1988; Sheppard & Meixner, 2003), representing thereby a substantial component of the total honey bee diversity. These 10 European subspecies have been grouped by morphological and molecular tools (Arias & Sheppard, 1996; Garnery *et al.*, 1992; Garnery *et al.*, 1993; Ruttner, 1988; Wallberg *et al.*, 2014; Whitfield *et al.*,

2006) into two evolutionary lineages: the M-lineage, in Western Europe, and the C-lineage, in Eastern Europe.

Subspecies-specific genetic footprints can still be identified in Europe (Garnery *et al.*, 1998; Miguel *et al.*, 2007; Oleksa *et al.*, 2011; Pinto *et al.*, 2012; Strange *et al.*, 2008; Uzunov *et al.*, 2014), in spite of centuries of beekeeping (Crane, 2013), although introgression and admixture events have also been detected in eastern (Muñoz *et al.*, 2009; Nedić *et al.*, 2014; Uzunov *et al.*, 2014) and western (Jensen *et al.*, 2005; Oleksa *et al.*, 2011; Pinto *et al.*, 2014; Soland-Reckeweg *et al.*, 2009) European populations. The M-lineage *A. m. mellifera* (Dark honey bee) has been recognized as the most threatened, with most of the threat due to introgression from the C-lineage (Jensen *et al.*, 2005; Pinto *et al.*, 2014; Soland-Reckeweg *et al.*, 2009). In addition to the documented intentional replacement of *A. m. mellifera* by *A. m. carnica* in Germany (Dreher, 1946; Maul & Hähnle, 1994), the increasing trade of commercial breeds (mainly C-lineage *A. m. carnica, A. m. ligustica* and the hybrid buckfast) is threatening the genetic integrity of the native *A. m. mellifera* as many beekeepers prefer using commercial as opposed to native honey bees.

Increasing awareness that native honey bee diversity represents a valuable asset for sustainable beekeeping is fuelling local breeding and conservation efforts across Europe. One of the earliest, and until recently the single conservation program enacted by law, is that implemented by the Danish Beekeepers Association and the Læsø Beekeepers Association on behalf of the Danish Government in 1993 and the European Union in 1998 (Jensen & Pedersen, 2005) to create a reserve and protect the *A. m. mellifera*. Following approval by the Scottish government of an order to protect the *A. m. mellifera* on the islands of Colonsay and Oronsay [The Bee Keeping (Colonsay and Oronsay) Order 2013], a second European reserve was recently created in the United Kingdom. Other *A. m. mellifera* conservation efforts, although not enacted by law, are underway in France, Holland, Norway, Switzerland, Ireland, and Belgium, among others (see the website "http://www.sicamm.org" run by the International Association for the Protection of the European Dark bee). The success or failure of all these efforts will be tightly linked to efforts that monitor the integrity of these protected populations.

Assessing introgression is an important activity in honey bee breeding programs, especially when conservation of native subspecies is a major concern. This activity requires molecular tools that are reliable, inexpensive and preferably automated. Previous studies have monitored introgression between the endemic *A. m. mellifera* and introduced C-lineage subspecies using

microsatellite and PCR-RFLP markers (Jensen *et al.*, 2005; Rortais *et al.*, 2011; Soland-Reckeweg *et al.*, 2009). However, with the publication of the honey bee genome (Weinstock *et al.*, 2006), development of single-nucleotide polymorphism (SNP) markers (Chávez-Galarza *et al.*, 2013; Whitfield *et al.*, 2006), and next generation sequencing becoming fast and affordable, particularly for a small genome as that of the honey bee (236 Mb), increasingly powerful tools are available to measure genomic ancestry and admixture levels occurring in both native and introduced honey bee populations (Harpur *et al.*, 2014; Harpur *et al.*, 2013; Wallberg *et al.*, 2014). However, the genomic approach is not always cost-effective and low quality and/or degraded DNA can be a handicap to using genomic re-sequencing. Alternatively, ancestry can be estimated using a subset of highly informative SNPs ranging in number from a few dozens to several hundreds. The selected SNPs, commonly known as Ancestry-Informative Markers (AIMs), are those that exhibit large allele frequency differences between populations. AIMs can be used for inferring geographic origin of individuals (Galanter *et al.*, 2012; Kosoy *et al.*, 2009; Rosenberg *et al.*, 2003), detecting illegal trade and translocation of animals (Frantz *et al.*, 2006), food authentication (Wilkinson *et al.*, 2012), for estimating overall admixture proportions efficiently and inexpensively (Galanter *et al.*, 2012; Parra *et al.*, 1998), among others. It is possible, using a panel of AIMs distributed throughout the genome, to estimate the relative ancestral proportions in admixed individuals, and infer the time since the admixture process (Falush *et al.*, 2003; Hoggart *et al.*, 2004).

The ability of an AIMs panel to measure ancestry is generally evaluated empirically, by examining its performance on a given set of samples for which ancestry is known (Pardo-Seco *et al.*, 2014). In this paper, we employed five analytical methods to select different combinations of SNPs to form five nested panels of 48-, 96-, 144-, 192- and 384-AIMs optimized to estimate admixture proportions of C-lineage (*A. m. ligustica* and *A. m. carnica*) into the M-lineage *A. m. mellifera*. This was done in two successive stages. In the first stage, we evaluated the performance of the five selection methods [Weir & Cockerham's F_{ST}, an F_{ST}-based outlier test, Delta, informativeness (In), and PCA] on a training dataset, in an effort to select AIMs and to rank them by decreasing level of informativeness. In the second stage, we tested the power of the reduced five designed panels and validated their performance on holdout and simulated sets, by comparing the admixture estimates produced by the panels with those produced by an initial dataset of 1183 SNPs.

Material and Methods

Samples, DNA Extraction and SNP Genotyping

A total of 113 honey bee haploid males were collected in 2010 and 2011 across the native range of *A. m. mellifera*, *A. m. ligustica* and *A. m. carnica* in Europe (see the sampling map Pinto *et al.*, 2014). The samples of *A. m. mellifera* (N=77) were collected from apiaries located in England (N=8), France (N=15), Belgium (N=3), Denmark (N=10), Holland (N=15), Switzerland (N=6), Scotland (N=10), and Norway (N=10) from protected and unprotected populations (Pinto *et al.*, 2014). Colonies of protected populations have been identified by morphological (B. Dahle, pers. comm.) and molecular tools (mtDNA $tRNA^{leu}$-cox2 and microsatellites (Bertrand *et al.*, 2015; Francis *et al.*, 2014; Jensen *et al.*, 2005; Soland-Reckeweg *et al.*, 2009) as the best representatives of *A. m. mellifera* and have therefore been integrated into conservation programs. To prevent C-lineage introgression and assure pure breeding, these colonies have been maintained in islands or in isolated mating stations. Despite careful management to protect the threatened *A. m. mellifera* from C-lineage introgression, a recent SNP survey detected variable, although generally low, levels of introgression in these protected populations (see Pinto *et al.*, 2014 for details). A reference collection of 36 samples representing C-lineage diversity was obtained from the natural range of *A. m. carnica* in Serbia (N=8) and Croatia (N=11) and from the natural range *A. m. ligustica* in Italy (N=17). The owners of all the sampled apiaries gave permission to collect honey bee individuals from the hives. In each location, samples were taken from the inner part of hives, placed into absolute ethanol and stored at −20 °C until molecular analysis.

Using a phenol/chloroform isoamyl alcohol (25:24:1) protocol (Sambrook & Russell, 1989), total DNA was extracted from the thorax of the 113 individuals, each representing a single colony. A total of 1536 SNP loci were genotyped for those individuals using Illumina's BeadArray Technology and the Illumina GoldenGate® Assay with a custom Oligo Pool Assay (Illumina, San Diego, CA, USA) following manufacturer's protocols. The Oligo Pool consisted of the 1536 SNPs, which included the 768 most informative SNPs (Whitfield *et al.*, 2006) and 768 newly developed SNPs employed by Chávez-Galarza *et al.* (2013). The 1536 SNP array was used previously to study diversity and introgression levels in populations of *A. m. mellifera* sampled across Western Europe (Pinto *et al.*, 2014) and to detect signatures of selection in the Iberian honey bee genome (Chávez-Galarza *et al.*, 2013). Genotype calling was performed using Illumina's GenomeStudio®

Data Analysis software. Of the initial 1536 SNPs, 353 did not meet the quality criteria for analysis and were therefore excluded from the dataset. The SNP filtering was as follows: 124 exhibited poorly separated intensity clusters or low signal intensity when visualized in the GenomeStudio software; 167 were monomorphic (defined by a cut-off criterion of > 0.98 for the most common allele, as in Chávez-Galarza *et al.* (2013) across all populations; 54 did not map in the honey bee genome assembly Amel_4.0; and 8 hit two different genomic positions (the first with 100% identity and the second with 96-98%) in the honey bee genome assembly Amel_4.0 during the mapping process using the 100 bp flanking sequence. Allele frequencies were calculated for each of the remaining 1183 bi-allelic SNPs (Table Sup. III-1) in each population using the program Plink (Purcell *et al.*, 2007).

Selection of AIMs

Five different methods were employed on the initial 1183 SNP dataset for estimating marker information content. The first method, which has been one of the most popular for selecting informative loci, was the pairwise F_{ST} of Weir & Cockerham (1984) as calculated at each locus using Genepop software (Raymond & Rousset, 2004). The second method was the F_{ST}–based outlier test developed by Foll & Gaggiotti (2008), which employs a Bayesian likelihood approach to detect loci deviating from neutral expectations (outliers). This outlier test was implemented in Bayescan 2.01 (Foll & Gaggiotti, 2008) using 20 pilot runs of 5,000 iterations (sample size of 5,000 and thinning interval of 10) and an additional burn-in of 50,000 iterations. The third method was based on the estimate of allele-frequency differential (Delta), which is one of the most straightforward ways to evaluate the information content of a SNP. For a bi-allelic marker, like a SNP, the Delta value is estimated as $|pA_i - pA_j|$, where pA_i and pA_j are the frequencies of allele A in the i^{th} and j^{th} populations, respectively. When more than two populations were analyzed, the Delta value for each SNP locus was estimated as the mean across all pairwise comparisons. The fourth method was the informativeness for assignment (I_n, natural logarithm of the number of populations) proposed by Rosenberg *et al.* (2003). I_n provides the amount of information gained about population assignment from observation of a single randomly chosen allele at a locus. This method assumes a uniform prior across K potential source populations for the origin of the allele. For a given set of populations, the minimum value of I_n (0) occurs when all alleles have equal frequencies in all populations whereas the maximum value (1) occurs when alleles are not shared

among populations. I_n was calculated using the software Infocalc available at http://www.stanford.edu/group/rosenberglab/infocalc.html. Finally, the fifth selection method was principal component analysis (PCA), which was performed using the PAST software (Hammer *et al.*, 2001). The first eight principal components were used to calculate the information content of each SNP following the approach of Paschou *et al.* (2007). The loadings for each SNP were squared and summed over the eight most significant principal components to produce an estimate of informativeness.

SNPs were ranked and panels of SNPs tested using reference populations and the Anderson's Simple Training and Holdout method to reduce the potential for upward bias, which is introduced when loci are ranked and assessed using the same individuals (Anderson, 2010). To that end, a total of 34 pure (*sensu* Soland-Reckeweg *et al.*, 2009) individuals of *A. m. mellifera*, previously identified in Pinto *et al.* (2014), and all reference individuals (17 *A. m. ligustica* and 19 *A. m. carnica*) were used for SNP ranking (training set=70) and the remaining 43 individuals of *A. m. mellifera* were reserved for panel testing (holdout set=113). To minimize the effect of clusters of populations on the selection of the AIMs (Ozerov *et al.*, 2013; Rosenberg *et al.*, 2003; Storer *et al.*, 2012), the five selection methods were tested using four training datasets. The first dataset consisted of 70 individuals: 34 pure *A. m. mellifera* and 36 C-lineage individuals, with no distinction between the *A. m. carnica* and *A. m. ligustica* subspecies (dataset I). The second dataset consisted of 51 individuals: 34 pure *A. m. mellifera* and 17 *A. m. ligustica* (dataset II). The third dataset consisted of 53 individuals: 34 pure *A. m. mellifera* and 19 *A. m. carnica* (dataset III). Finally, the fourth dataset consisted of 70 individuals: 34 pure *A. m. mellifera*, 17 *A. m. ligustica* and 19 of *A. m. carnica* (dataset IV).

Ranking of SNPs

The five selection methods were implemented on the four training datasets producing a total of 20 information content values for each of the 1183 SNPs. These values were ranked and analyzed individually and then were averaged in two steps to obtain a single global value per SNP. In the first step the information content values were averaged across the four training datasets for each of the five selection methods. In the second step the information content values produced by each selection method were converted into a 0-1 scale and then averaged to obtain a global score for each of the 1183 SNPs. After standardizing the values produced by the five selection methods, the

global ranking was obtained for the 1183 SNPs using the global score. Given that linked loci yield redundant information, having therefore similar resolving power, markers were excluded if they were within a predefined genetic distance (< 1 cM) of higher ranking selected SNPs. The genetic distance of the remaining SNPs ranged from 1.01 to 24.25 cM with a mean of 4.64 cM. Prior to obtaining the global score for each SNP, pairwise associations between information content values produced by the five methods and between the four training datasets were calculated using the Spearman`s rank correlation coefficient, in order to compare the five selection methods and examine the effect of clusters of populations.

Panel Testing

Five panels of 48-, 96-, 144-, 192- and 384-SNPs (sets defined by multiplex sizes of commercial assays) were designed from the top-ranked SNPs. These nested panels were tested against a holdout set and a simulated set to obtain the admixture proportions estimated by each SNP panel. The holdout set (113 individuals) consisted of 34 pure individuals plus 43 reserved individuals of *A. m. mellifera* and the reference *A. m. ligustica* (17 individuals) and *A. m. carnica* (19 individuals), as described above. The simulated set (1,000 individuals) was generated with the program ONCOR (Kalinowski *et al.*, 2007) using the function "simulate a single mixture". Ten populations, each with 100 simulated genotypes, were simulated using different levels of introgression (0, 1, 5, 10, 20, 30, 40, 50, 75, and 90%).

Two approaches were used to validate the five reduced AIMs panels. First, a PCA was performed with SNPs in each AIMs panel on the holdout set using the software PAST to generate two-dimensional PCA and to visualize the stability of population assignment produced by the panels. Second, ancestry and admixture was analyzed. Admixture proportions were estimated with SNPs in each AIMs panel for the holdout and simulated sets using a model-based maximum likelihood estimation of individual ancestries implemented in the software Admixture v1.23 (Alexander *et al.*, 2009). Coancestry spanning 1-6 populations (K=1-6, using the default termination criterion that stops the runs when the log-likelihood increases by less than $\varepsilon < 0.0001$ between iterations) was explored for each AIMs panel and the optimal K was identified with the inferred number of populations producing the lowest cross-validation error (CV) during the clustering analysis.

The performance of each reduced panel was examined using different approaches. First, the pairwise differences between admixture proportions inferred from the initial 1183 SNP dataset and the five panels were tested using a Mann-Whitney test. Second, the precision of each panel was tested against the initial 1183 SNP dataset by calculating linear regression coefficients (r^2) and the standard deviations of the differences between admixture proportions. Finally, the accuracy of the reduced panels was estimated via percentage of absolute error of admixture estimates obtained with the five panels in relation to the initial 1183 SNP dataset.

Results

Identification and Ranking of AIMs

The majority of the 1183 SNPs assessed in this study using five selection methods (pairwise Weir & Cockerham's F_{ST}, F_{ST}-based outlier test, Delta, I_n and PCA) contain high levels of information content (Figure III-1, Table Sup. III-2), facilitating the design of reduced panels for genetic identification and introgression analysis in the Dark honey bee, *A. m. mellifera*. The distribution of frequency histograms and percentiles of genetic information content of the 1183 SNPs estimated by each selection method and training dataset are shown in Figure III-1. The 50^{th} percentile ranges of the four training datasets were 0.6974-0.7712, 0.5459-0.6362, 0.5532-0.7601, 0.3345-0.3583 and 0.0038-0.0040 for the Weir & Cockerham's F_{ST}, F_{ST}-based outlier test, Delta, I_n and PCA, respectively, indicating a high level of information content for most SNPs and a similar pattern among the four training datasets (Figure III-1).

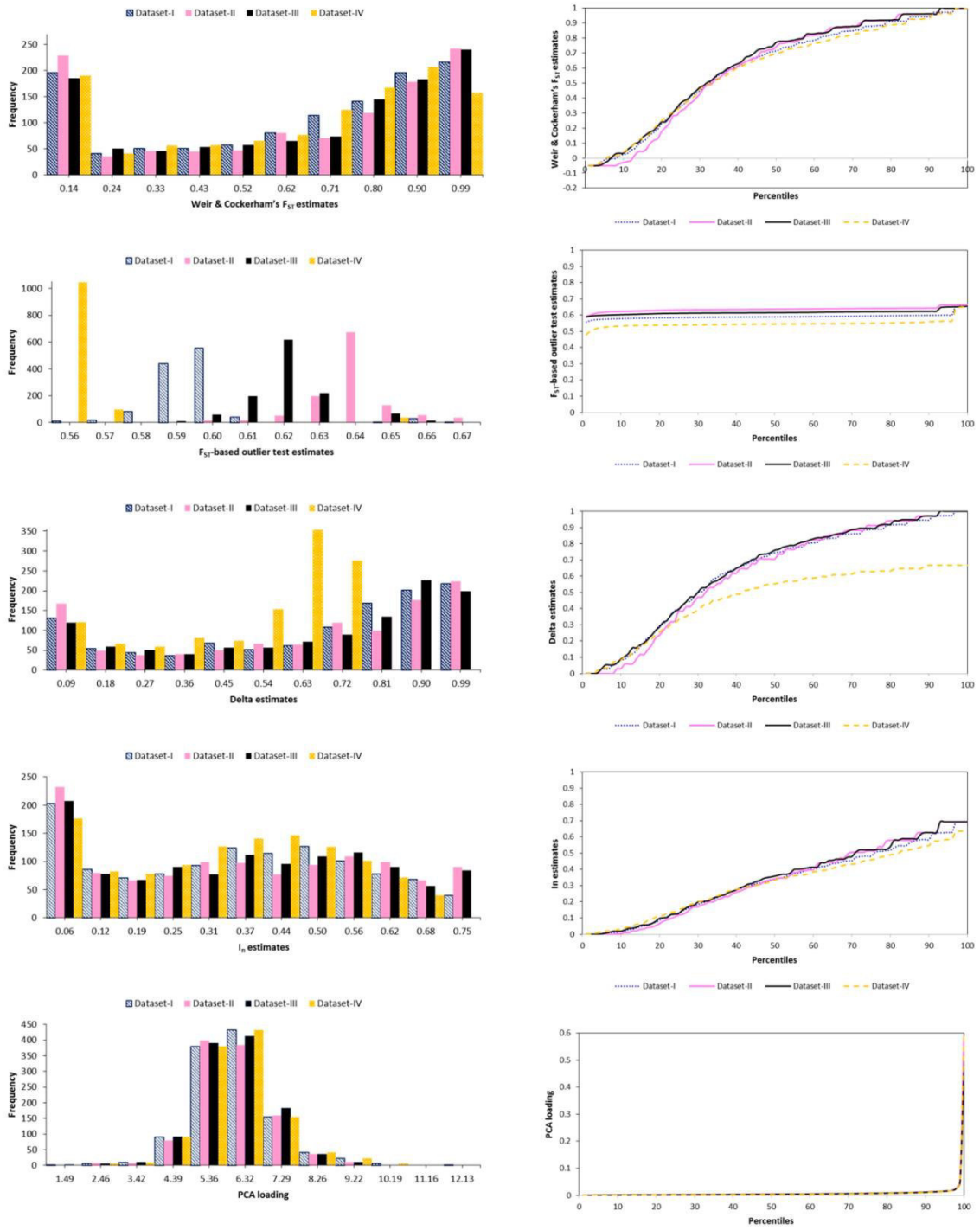


Figure III-1 - Frequency histograms and percentiles of the estimates of genetic information contained in the initial 1183 SNP dataset. Information content produced by the five selection methods (pairwise Weir & Cockerham's F_{ST}, F_{ST}-based outlier test, Delta, I_n and PCA) is shown for the four training datasets (I, II, III and IV).

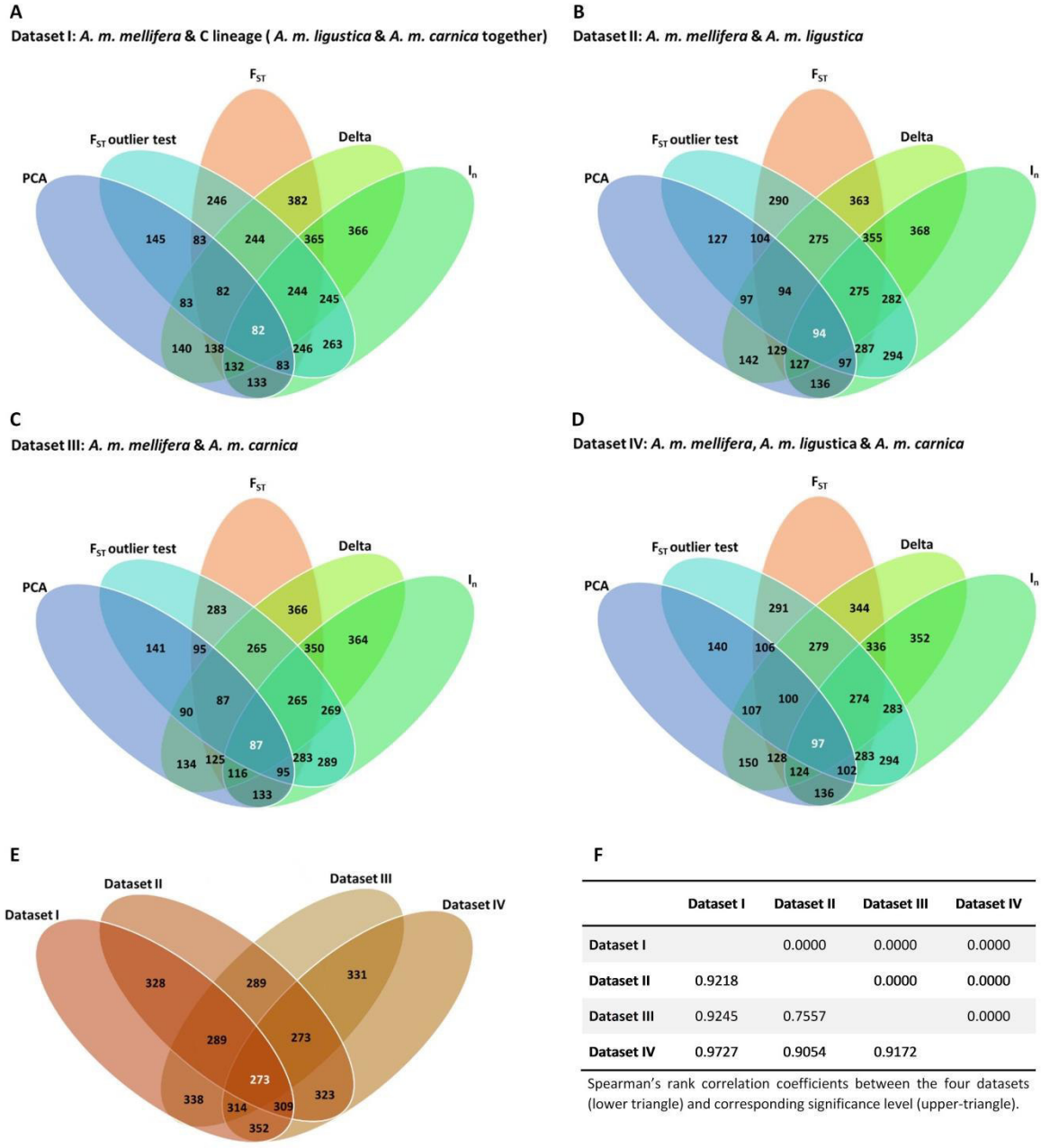
The level of similarity (Spearman's rank correlation, r_s) between the different estimates of genetic information content produced by the five selection methods across the four training datasets is shown in . The highest correlation values were observed for Weir & Cockerham's F_{ST}, Delta and I_n ($0.7648 \leq r_s \leq 0.9985$, $P < 0.001$) whereas a moderate correlation was detected between the F_{ST}-based outlier test and Weir & Cockerham's F_{ST}, Delta and I_n ($0.2864 \leq r_s \leq 0.6592$, $P < 0.001$). The lowest correlations were observed between PCA and the other four methods ($-0.2228 \leq r_s \leq 0.1025$, $0.000 \leq P \leq 0.9412$). Regarding the four training datasets (Figure III-2), high correlation values were observed across selection methods ($0.7557 \leq r_s \leq 0.9727$, $P < 0.001$).

Using an information content cutoff value ≥ 0.25, which indicates very great genetic differentiation (Wright, 1978), a total of 627 AIMs were identified by the methods of Weir & Cockerham's F_{ST}, Delta, I_n, and F_{ST}-based outlier test. Of these, the top-ranked 384 AIMs were selected using the five methods and the four training datasets. The extent of overlap of the 384 AIMs across the five selection methods and the four training datasets is shown in Figure III-2. Overlap between any two methods and across datasets ranged between 382 (Weir & Cockerham's F_{ST} and Delta for dataset I; Figure III-2A) and 134 (Delta and PCA for dataset III; Figure III-2C). The number of AIMs that were simultaneously selected by the five methods was lower, ranging from 82 (dataset I; Figure III-2A) to 97 (dataset IV; Figure III-2D). A substantially higher amount of overlap (273 AIMs; Figure III-2E), supported by high correlation values ($r_s \geq 0.7557$, $P < 0.001$; Figure III-2F), was observed across the four training datasets, suggesting that the different population groupings have a small effect on the AIMs ranking. The global ranking of the 384 AIMs was used to design reduced panels of 192-, 144-, 96-, and 48 that included SNPs with the highest respective global scores. The performance of these reduced panels was subsequently assessed using the holdout and simulated sets.

Table III 1 - Comparison of selection methods and training datasets.

		Dataset I: *A. m. mellifera* & C lineage (*A. m. ligustica* & *A. m. carnica together)*					Dataset II: *A. m. mellifera* & *A. m. ligustica*					Dataset III: *A. m. mellifera* & *A. m. carnica*					Dataset IV: *A. m. mellifera A. m. ligustica & A. m. carnica*				
		F_{ST}	Delta	I_n	PCA	F_{ST} outlier test	F_{ST}	Delta	I_n	PCA	F_{ST} outlier test	F_{ST}	Delta	I_n	PCA	F_{ST} outlier test	F_{ST}	Delta	I_n	PCA	F_{ST} outlier test
Dataset I	F_{ST}		0.0000	0.0000	0.7134	0.0000	0.0000	0.0000	0.0000	0.0620	0.0000	0.0000	0.0000	0.0000	0.6130	0.0000	0.0000	0.0000	0.0000	0.7134	0.0000
	Delta	0.9985		0.0000	0.6139	0.0000	0.0000	0.0000	0.0000	0.0439	0.0000	0.0000	0.0000	0.0000	0.6772	0.0000	0.0000	0.0000	0.0000	0.6139	0.0000
	I_n	0.9977	0.9957		0.7284	0.0000	0.0000	0.0000	0.0000	0.0753	0.0000	0.0000	0.0000	0.0000	0.4980	0.0000	0.0000	0.0000	0.0000	0.7284	0.0000
	PCA	0.0107	0.0147	0.0101		0.0000	0.9412	0.8104	0.7135	0.0000	0.0000	0.0390	0.0241	0.0357	0.0000	0.0711	0.4901	0.1065	0.5267	0.0000	0.0001
	F_{ST} outlier test	0.5974	0.5778	0.6280	-0.1195		0.0000	0.0000	0.0000	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Dataset II	F_{ST}	0.9364	0.9370	0.9392	-0.0021	0.5666		0.0000	0.0000	0.3782	0.0000	0.0000	0.0000	0.0000	0.8227	0.0000	0.0000	0.0000	0.0000	0.9412	0.0000
	Delta	0.9338	0.9357	0.9313	-0.0070	0.5205	0.9906		0.0000	0.3980	0.0000	0.0000	0.0000	0.0000	0.7760	0.0000	0.0000	0.0000	0.0000	0.8104	0.0000
	I_n	0.9329	0.9324	0.9342	-0.0107	0.5581	0.9923	0.9960		0.5187	0.0000	0.0000	0.0000	0.0000	0.8981	0.0000	0.0000	0.0000	0.0000	0.7135	0.0000
	PCA	0.0543	0.0586	0.0517	0.8545	-0.1074	0.0256	0.0246	0.0188		0.0000	0.0017	0.0004	0.0008	0.0000	0.1510	0.0986	0.0426	0.1329	0.0000	0.0000
	F_{ST} outlier test	0.4404	0.4235	0.4637	-0.2018	0.8128	0.4998	0.4732	0.5124	-0.2228		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Dataset III	F_{ST}	0.9328	0.9317	0.9353	0.0600	0.5751	0.7965	0.7705	0.7719	0.0909	0.3301		0.0000	0.0000	0.2647	0.0000	0.0000	0.0000	0.0000	0.0390	0.0000
	Delta	0.9347	0.9353	0.9327	0.0656	0.5337	0.7825	0.7690	0.7663	0.1025	0.2864	0.9925		0.0000	0.4242	0.0000	0.0000	0.0000	0.0000	0.0241	0.0000
	I_n	0.9331	0.9312	0.9349	0.0611	0.5742	0.7829	0.7648	0.7656	0.0969	0.3150	0.9943	0.9960		0.3212	0.0000	0.0000	0.0000	0.0000	0.0357	0.0000
	PCA	-0.0147	-0.0121	-0.0197	0.6450	-0.1333	-0.0065	0.0083	0.0037	0.5534	-0.1435	-0.0325	-0.0233	-0.0289		0.0000	0.3690	0.3671	0.4471	0.0000	0.0000
	F_{ST} outlier test	0.5524	0.5361	0.5793	-0.0525	0.8862	0.4451	0.3910	0.4189	-0.0418	0.6063	0.6337	0.6006	0.6444	-0.1440		0.0000	0.0000	0.0000	0.0711	0.0000
Dataset IV	F_{ST}	0.9883	0.9862	0.9875	0.0201	0.6028	0.9301	0.9216	0.9222	0.0480	0.4567	0.9411	0.9387	0.9384	-0.0261	0.5771		0.0000	0.0000	0.4901	0.0000
	Delta	0.9459	0.9485	0.9474	0.0470	0.5487	0.9177	0.9022	0.9014	0.0590	0.4237	0.9241	0.9169	0.9152	-0.0262	0.5475	0.9683		0.0000	0.1065	0.0000
	I_n	0.9825	0.9798	0.9832	0.0184	0.6081	0.9313	0.9277	0.9308	0.0437	0.4733	0.9273	0.9279	0.9300	-0.0221	0.5798	0.9948	0.9710		0.5267	0.0000
	PCA	0.0107	0.0147	0.0101	1.0000	-0.1195	-0.0021	-0.0070	-0.0107	0.8545	-0.2018	0.0600	0.0656	0.0611	0.6450	-0.0525	0.0201	0.0470	0.0184		0.0001
	F_{ST} outlier test	0.6043	0.5867	0.6267	-0.1154	0.9119	0.5549	0.5292	0.5634	-0.1250	0.8250	0.5819	0.5589	0.5943	-0.1376	0.8758	0.6344	0.6104	0.6592	-0.1154	

Spearman's rank correlation coefficients (lower triangle), and corresponding P-values (upper triangle), between information content estimates produced by the five selection methods (Weir & Cockerham's FST, Delta, informativeness (In), PCA, FST-based outlier test) using the four training datasets (I, II, III and IV).



F

	Dataset I	Dataset II	Dataset III	Dataset IV
Dataset I		0.0000	0.0000	0.0000
Dataset II	0.9218		0.0000	0.0000
Dataset III	0.9245	0.7557		0.0000
Dataset IV	0.9727	0.9054	0.9172	

Spearman's rank correlation coefficients between the four datasets (lower triangle) and corresponding significance level (upper-triangle).

Figure III-2- (A-E) Venn diagrams showing the extent of overlap of the top-ranked 384 AIMs. (A-D) Overlap among the five selection methods (pairwise Weir & Cockerham's F_{ST}, F_{ST}-based outlier test, Delta, I_n and PCA) and the four training datasets (I, II, III and IV). (E) Overlap among the four training datasets, after averaging the information content obtained with the five selection methods, and (F) corresponding Spearman's rank correlation coefficients.

Validation of the AIMs Panels

The performance of the five AIMs panels (48-, 96-, 144-, 192-, 384-AIMs) was first validated by using PCA to produce a visual summary of the observed genetic variation carried by the holdout set (Figure III-3). The overall diversity pattern is characterized by the presence of two distinct clusters, which are coincidental with the M and C evolutionary lineages. This pattern was captured by every single AIMs panel, although a greater dispersion was observed for the smaller panels. Additionally, the panels with less than 192 AIMs were unable to distinguish the two C-lineage subspecies, *A.m. ligustica* and *A.m. carnica*, which were clearly identified by the initial 1183 SNPs and, to a lesser degree, by the 384-AIMs panel.

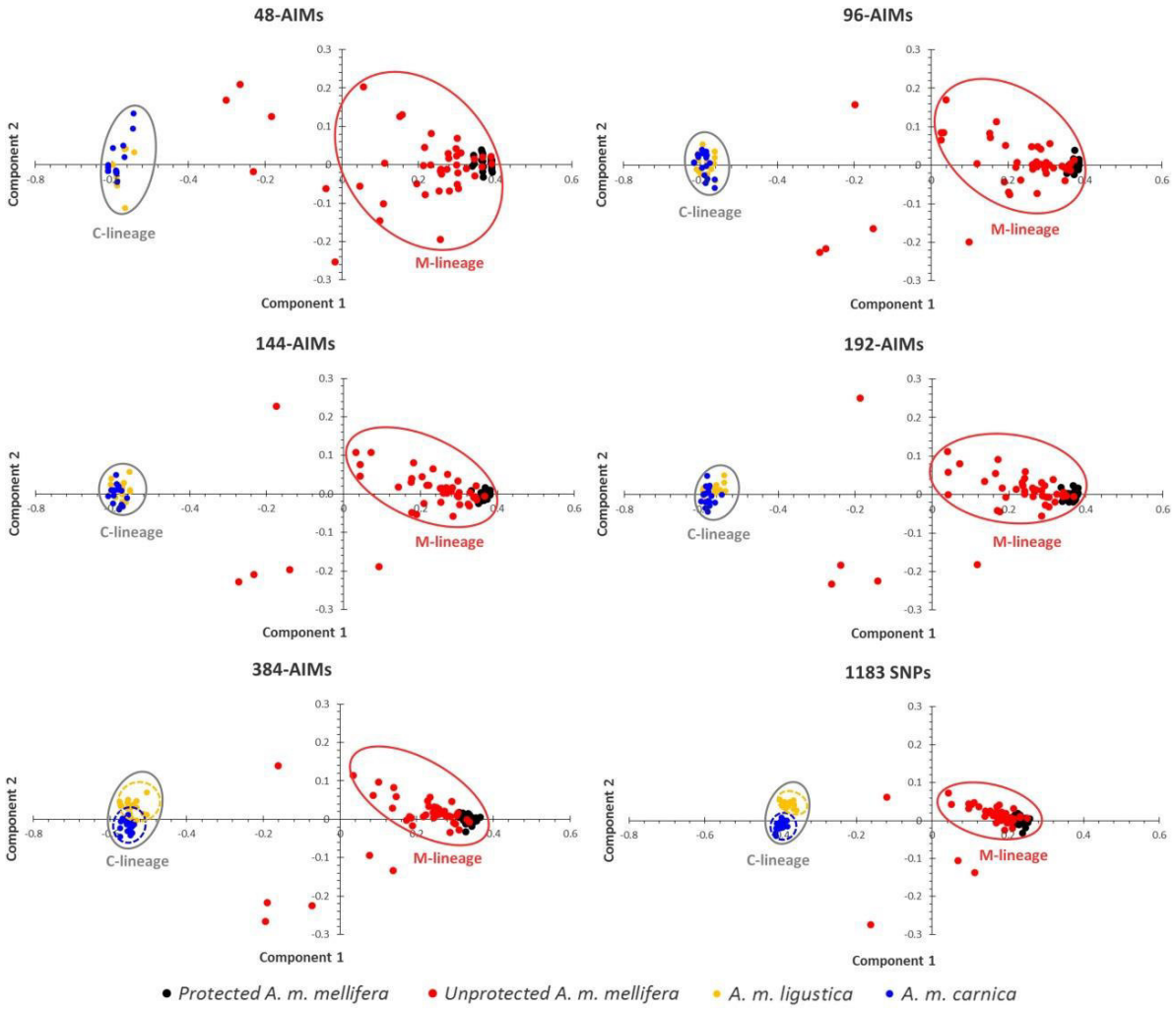


Figure III-3 – Principal components analysis. Plots obtained for the holdout set using the five AIMs panels (48-, 96-, 144-, 192-, 384-AIMs) and the initial 1183 SNP dataset.

Ancestry and admixture analyses based on admixture estimates confirm the overall pattern captured by the PCA (Table Sup. III-3 and Figure Sup. III-1). At the optimal K=2 (inferred by the initial 1183 SNP dataset and the five AIMs panel), the two clusters corresponded to the C and M-

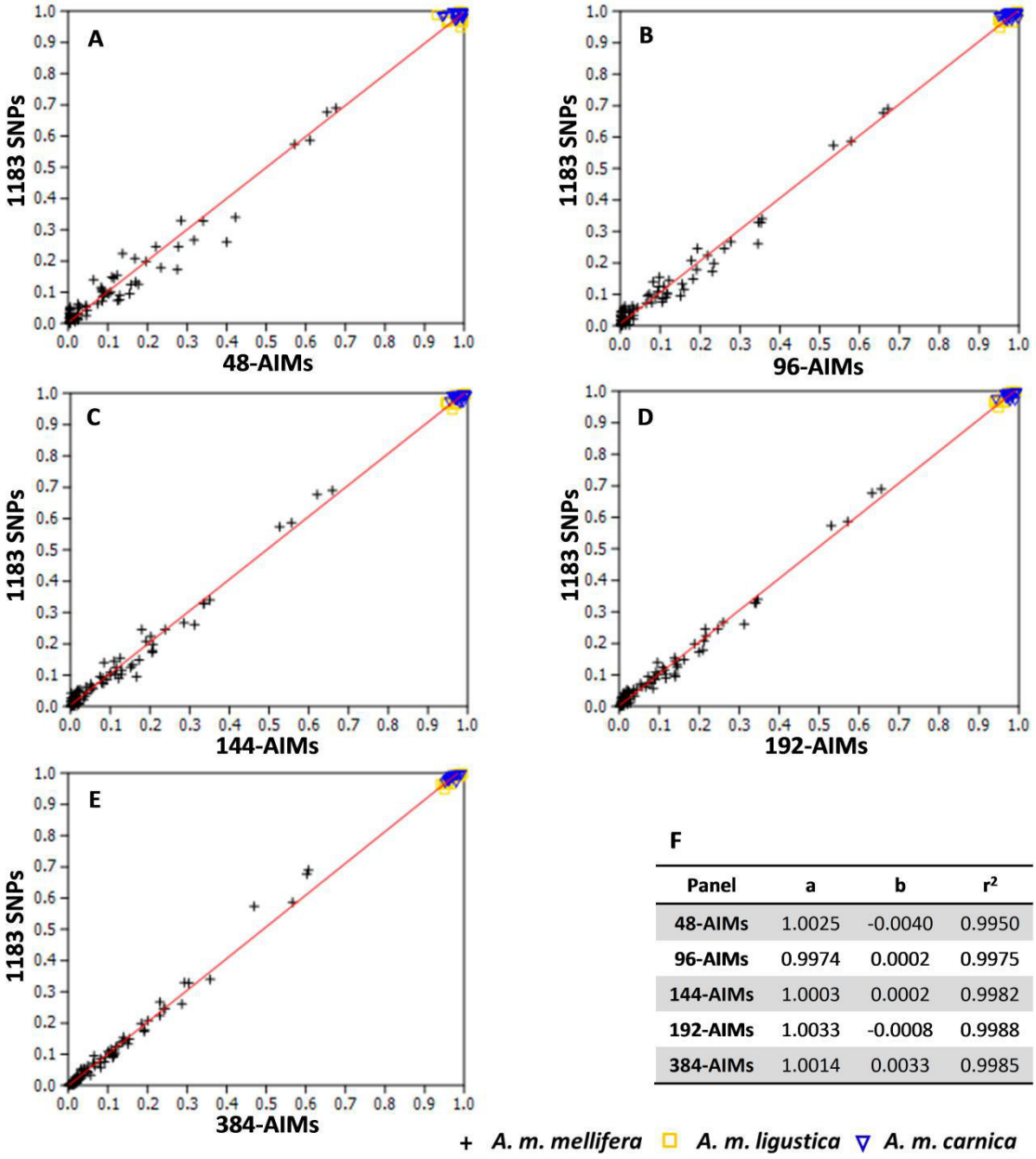
lineages. However, C-lineage individuals formed a more homogeneous cluster than those of the M-lineage individuals. While membership proportions in the C-lineage cluster were greater than 95% for the five AIMs panels, the M-lineage cluster comprised 13 (384-AIMs and 1183 SNPs), 14 (48- and 192-AIMs) and 15 (96- and 144-AIMs) individuals with membership proportions lower than 85%, a pattern that was already evident in the PCA plots.

The introgression levels exhibited by individuals of the M-lineage cluster were significantly higher (Student's t-test, $P < 0.001$) in unprotected (13.76-15.18%, with 1183 SNPs and 48 AIMs, respectively) than in protected individuals (0.08-0.52%, with 96 AIMs and 1183 SNPs, respectively) for any AIMs panel. The overall estimates of C-lineage introgression into *A. m. mellifera* varied with the panel (8.4, 7.9, 7.8, 7.9, 7.5 and 7.7% with 48-, 96-, 144-, 192-, 384-AIMs and 1183 SNPs, respectively), although the differences were not statistically significant (Mann-Whitney test, $0.8225 \leq P \leq 0.9983$; Table Sup. III-4).

In addition to the admixture analyses using the holdout set, the AIMs panels were further validated using a simulated set of 10 different levels of C-lineage introgression (0, 1, 5, 10, 20, 30, 40, 50, 75, and 90%). As for the analyses with the holdout set, the simulated set produced two clusters corresponding to M and C lineages with no significant differences in admixture proportions between the different AIMs panels and the initial 1183 SNP dataset (Mann-Whitney test, $P \geq 0.2313$; Table Sup. III-5).

Assignment's precision and accuracy

The power of the reduced AIMs panels in identifying *A. m. mellifera* and estimating admixture proportions was evaluated on the holdout set. Estimates of C-lineage introgression into *A. m. mellifera* inferred from the five panels were greatly concordant with those inferred from the initial 1183 SNP dataset, as indicated by the high correlation values ($r \geq 0.997$; Figure III-4). Despite the high correlations obtained for each comparison, the error rate in admixture estimates, which is very low for all the panels (0.0012–0.0042 with the simulated set and 0.4–1.3 with the holdout set), does increase as the size of the panel decreases (Figure Sup. III-2). Nevertheless, the reduced AIMs panels provide good precision in estimating admixture proportions.



F

Panel	a	b	r^2
48-AIMs	1.0025	-0.0040	0.9950
96-AIMs	0.9974	0.0002	0.9975
144-AIMs	1.0003	0.0002	0.9982
192-AIMs	1.0033	-0.0008	0.9988
384-AIMs	1.0014	0.0033	0.9985

Figure III-4 – Linear regression. (A-E) Plots between admixture proportions inferred from the initial 1183 SNP dataset and those inferred from the five AIMs panels (48-, 96-, 144-, 192-, 384-AIMs) using individuals of the holdout set. (F) Parameters and coefficients for each AIMs panel.

As another assessment of the performance of the panels, the accuracy was calculated via absolute error. The success of assignment of the 113 individual genotypes of the holdout set to genetic origin and level of admixture inferred from the different AIMs panels is shown in Figure III-5. The average percentage of correct assignment was high varying from 98.2, 98.8, 99.0, 99.2 to 99.4% for the 48-, 96-, 144-, 192- and 384-AIMs panels, respectively. The chosen AIMs panels accurately distinguish M/C admixture, therefore these results suggest that a small number of AIMs

are sufficient to identify *A. m. mellifera* and estimate introgression from C-lineage colonies with great accuracy.

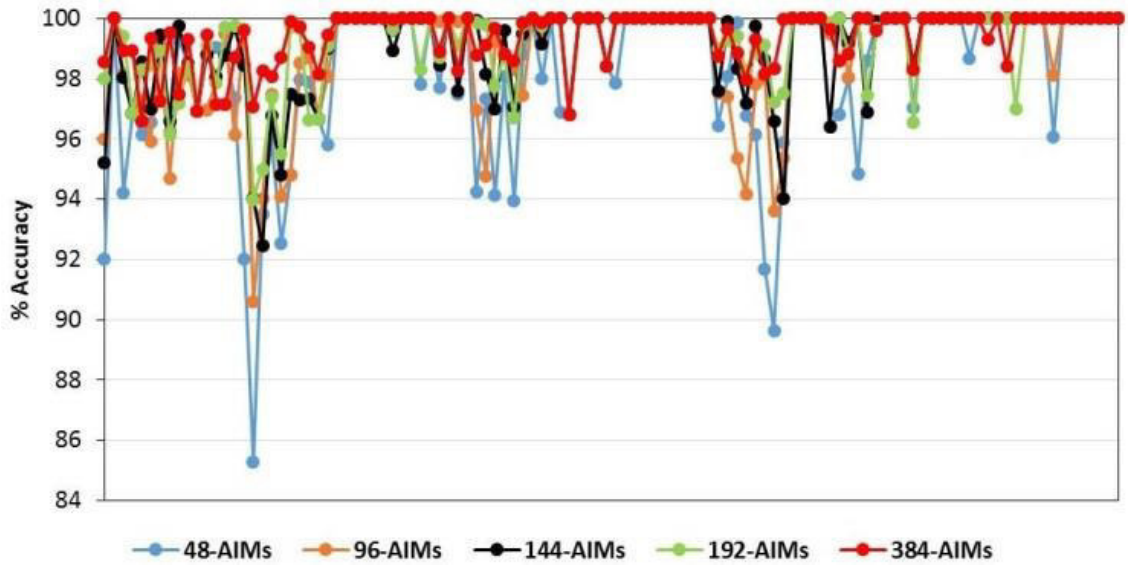


Figure III-5 – Assignment accuracy. Percentage obtained with the five AIMs panels (48-, 96-, 144-, 192-, 384-AIMs) for each of the 113 individuals of the holdout set.

Discussion

The recognition that native honey bee genetic diversity is fundamental for sustainable beekeeping and for facing the challenges of a rapidly changing world (e.g. climate change, novel diseases and parasites) is stimulating implementation of conservation programs across Europe in an attempt to recover and protect *A. m. mellifera*, which is the European honey bee subspecies with the widest natural range (Ruttner, 1988), and at the same time the most threatened by introgression (Jensen *et al.*, 2005; Pinto *et al.*, 2014). The need of a reliable, high-throughput, and cost-effective tool for identifying candidate *A. m. mellifera* colonies targeted for conservation, a crucial step when managing conservatoires, motivated the design of reduced AIMs panels containing the most informative SNPs to verify ancestry and introgression from C-lineage subspecies. In this study we developed, validated and tested the first reduced AIMs panels for honey bees. Our results provide strong confidence in a panel of 384 AIMs and show that even smaller subsets of 192-, 144-, 96- and 48-AIMs are able to identify ancestry and estimate introgression with great accuracy. These reduced panels promise to be a useful tool for routine identification of *A. m. mellifera* colonies maintained in the breeding populations of conservation programs.

The AIMs included in the five reduced panels were simultaneously selected by pairwise Weir & Cockerham's F_{ST}, F_{ST}-based outlier test, Delta, I_n and PCA, in order to balance out the limitations of each individual method (Paschou *et al.*, 2007; Rosenberg *et al.*, 2003; Wilkinson *et al.*, 2011). These selection methods have proved to be powerful, although with varying performances, in identifying population informative markers in a wide range of organisms (Galanter *et al.*, 2012; Ozerov *et al.*, 2013; Paschou *et al.*, 2007; Storer *et al.*, 2012; Wilkinson *et al.*, 2011). A great extent of overlap of top-ranked AIMs was obtained for the five selection methods, especially for pairwise Weir & Cockerham's F_{ST}, Delta, and I_n suggesting that they capture the same information. Nonetheless, the smaller panels (48-, 96-, 144-, 192-AIMs) did not necessarily include all AIMs simultaneously detected by the five methods as the global ranking depended on the average score. High pairwise correlation values were obtained for Weir & Cockerham's F_{ST}, Delta and I_n but not for PCA, as found by Wilkinson *et al.*, (2011). PCA has been recommended for ranking markers because it has the advantage of generating an overall estimate for a single SNP locus whereas the other methods require estimate of an average from pairwise calculations when the number of populations is greater than two (Paschou *et al.*, 2007).

The five reduced panels tested with the holdout and simulated sets performed virtually as well as the initial 1183 SNP dataset, as revealed by the strong correlations obtained between admixture estimates and low associated error rates. The assignment power was high across the five panels with average values of correct assignment varying between 98.2 and 99.4%, although the accuracy decreased slightly with panel size. Nonetheless, even the 48-AIMs panel exhibited high accuracy levels, which is not surprising as it includes the AIMs with the greatest resolution power. Studies on other organisms have also found good performances with panels of similar sizes (Galanter *et al.*, 2012; Storer *et al.*, 2012; Wilkinson *et al.*, 2012; Wilkinson *et al.*, 2011), detecting sharp drops in accuracy for a number of SNPs below 25 (Storer *et al.*, 2012; Wilkinson *et al.*, 2012).

Evaluation of different combinations of the focal *A. m. mellifera* and the two most common sources of foreign genes, *A. m. ligustica* and *A. m. carnica,* revealed a negligible effect of population groupings on the AIMs ranking. These results suggest that the designed panels are suited for identifying and assessing introgression of *A. m. ligustica*, *A. m. carnica* or both into *A. m. mellifera*. While these panels will possibly perform well in the presence of other C-lineage subspecies, more complex combinations that include sources of different evolutionary lineages will

require further testing and, most likely, new panels developed from broader baseline datasets. Additionally, it should be noted, that these reduced panels are not suitable for standard population genetic analyses, including determining allelic diversity or measuring isolation by distance, genetic drift or bottleneck effect. The bias introduced through selection for markers that segregate among target populations would seriously compromise these calculations (Albrechtsen *et al.*, 2010; Clark *et al.*, 2005).

Ancestry identification of honey bee subspecies is undergoing steady development (reviewed Meixner *et al.*, 2013) from classical morphometry, analysis of allozymes, mitochondrial DNA, nuclear microsatellites, and now SNP tools. Because researchers must balance the cost of genotyping many samples versus many loci, herein we developed five nested reduced panels that include AIMs with the highest resolution power for discriminating subspecies of the divergent M and C evolutionary lineages. While the 384-AIMs panel is also capable of discriminating the C-lineage *A. m. ligustica* and *A. m. carnica*, for estimating C-lineage introgression into *A. m. mellifera* we recommend using the 96-AIMs panel because it is accurate; and high-throughput 96-plex genotyping assays can be outsourced at an affordable cost ($8 900 for 480 samples), representing a saving of 92.4% when compared with the 1536-plex assay ($116 800 for 480 samples).

In conclusion, the proposed AIMs panels can be actively used as a tool in conservation management of *A. m. mellifera* populations that suffer from hybridization and introgression with the most commonly introduced and beekeepers' preferred *A. m. ligustica* and *A. m. carnica* subspecies. This can be an important advance because the current European regulation on organic beekeeping states that "preference shall be given to the use of European breeds of *Apis mellifera* and their local ecotypes" and several conservation programs have been undertaken in Europe (reviewed by De la Rúa *et al.*, 2009). The use of these panels will apply well to monitoring, management and conservation programs of *A. m. mellifera* in Western Europe, which usually require high-sample throughput, and will be a resource for the honey bee community to obtain accurate genetic information at reduced costs.

Acknowledgments

We are deeply grateful to Andrew Abrahams, Bjørn Dahle, Gabriele Soland-Reckeweg, Gilles Fert, Lionel Garnery, Norman Carreck, Pilar de la Rúa, Raffaele Dall'Olio, and Romee Van der Zee for providing honey bee samples. DNA extractions and SNP genotyping were performed by Colette

Abbey, with support from the TAMU Institute of Genomic Science and Society. An earlier version of the manuscript was improved by the constructive comments made by two anonymous reviewers.

References

Albrechtsen, A., Nielsen, F. C., & Nielsen, R. (2010). Ascertainment biases in SNP chips affect measures of population divergence. *Molecular Biology and Evolution, 27*(11), 2534-2547. doi: 10.1093/molbev/msq148

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research, 19*(9), 1655-1664. doi: 10.1101/gr.094052.109

Allendorf, F. W., & Luikart, G. (2007). *Conservation and the genetics of populations* (Vol. 2007).

Anderson, E. (2010). Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. *Molecular Ecology Resources, 10*(4), 701-710.

Arias, M., & Sheppard, W. (1996). Molecular Phylogenetics of Honey Bee Subspecies (*Apis mellifera L.*) Inferred from Mitochondrial DNA Sequence. *Molecular phylogenetics and evolution, 5*(3), 557-566.

Bertrand, B., Alburaki, M., Legout, H., Moulin, S., Mougel, F., & Garnery, L. (2015). MtDNA COI-COII marker and drone congregation area: An efficient method to establish and monitor honeybee (*Apis mellifera L.*) conservation centres. *Molecular ecology resources, 15*(3), 673-683.

Büchler, R., Costa, C., Hatjina, F., Andonov, S., Meixner, M. D., Conte, Y. L., Uzunov, A., Berg, S., Bienkowska, M., & Bouga, M. (2014). The influence of genetic origin and its interaction with environmental effects on the survival of *Apis mellifera L.* colonies in Europe. *Journal of Apicultural Research, 53*(2), 205-214.

Chávez-Galarza, J., Henriques, D., Johnston, J. S., Azevedo, J. C., Patton, J. C., Muñoz, I., De la Rúa, P., & Pinto, M. A. (2013). Signatures of selection in the Iberian honey bee (*Apis mellifera iberiensis*) revealed by a genome scan analysis of single nucleotide polymorphisms. *Molecular Ecology, 22*(23), 5890-5907.

Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H., & Nielsen, R. (2005). Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research, 15*(11), 1496-1502. doi: 10.1101/gr.4107905

Crane, E. E. (2013). *The world history of beekeeping and honey hunting*: Routledge.

De la Rúa, P., Jaffé, R., Dall'Olio, R., Muñoz, I., & Serrano, J. (2009). Biodiversity, conservation and current threats to European honeybees. *Apidologie, 40*(3), 263-284. doi: 10.1051/apido/2009027

Dowling, T. E., & Secor, C. L. (1997). The role of hybridization and introgression in the diversification of animals. *Annual review of Ecology and Systematics, 28*(1), 593-619.

Dreher, K. (1946). Gedanken zum Neuaufbau des Zuchtwesens. *Die Hessische Biene, 81*, 62-64.

Engel, M. S. (1999). The taxonomy of recent and fossil honey bees (Hymenoptera: Apidae; *Apis*).

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics, 164*(4), 1567-1587.

Foll, M., & Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics, 180*(2), 977-993.

Francis, R. M., Kryger, P., Meixner, M., Bouga, M., Ivanova, E., Andonov, S., Berg, S., Bienkowska, M., Büchler, R., & Charistos, L. (2014). The genetic origin of honey bee colonies used in the COLOSS Genotype-Environment Interactions Experiment: a comparison of methods. *Journal of Apicultural Research, 53*(2), 188-204.

Frantz, A., Pourtois, J. T., Heuertz, M., Schley, L., Flamand, M.-C., Krier, A., Bertouille, S., Chaumont, F., & Burke, T. (2006). Genetic structure and assignment tests demonstrate illegal translocation of red deer (*Cervus elaphus*) into a continuous population. *Molecular Ecology, 15*(11), 3191-3203.

Galanter, J. M., Fernandez-Lopez, J. C., Gignoux, C. R., Barnholtz-Sloan, J., Fernandez-Rozadilla, C., Via, M., Hidalgo-Miranda, A., Contreras, A. V., Figueroa, L. U., & Raska, P. (2012). Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. *PLoS genetics, 8*(3), e1002554.

Garnery, L., Cornuet, J. M., & Solignac, M. (1992). Evolutionary history of the honey bee *Apis mellifera* inferred from mitochondrial DNA analysis. *Molecular ecology, 1*(3), 145-154.

Garnery, L., Franck, P., Baudry, E., Vautrin, D., Cornuet, J., & Solignac, M. (1998). Genetic diversity of the west European honey bee (*Apis mellifera mellifera* and *A. m. iberica*) II. Microsatellite loci. *Genetics Selection Evolution, 30*(1), S49.

Garnery, L., Solignac, M., Celebrano, G., & Cornuet, J.-M. (1993). A simple test using restricted PCR-amplified mitochondrial DNA to study the genetic structure of *Apis mellifera L. Experientia, 49*(11), 1016-1021.

Hammer, Ø., Harper, D., & Ryan, P. (2001). PAST: Paleontological Statistics Software Package for Education and Data Analysis–Palaeontol. Electron. 4: 9pp.

Harpur, B. A., Kent, C. F., Molodtsova, D., Lebon, J. M., Alqarni, A. S., Owayss, A. A., & Zayed, A. (2014). Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. *Proceedings of the National Academy of Sciences, 111*(7), 2614-2619.

Harpur, B. A., Minaei, S., Kent, C. F., & Zayed, A. (2013). Admixture increases diversity in managed honey bees: Reply to De la Rúa et al. 2013. *Molecular ecology, 22*(12), 3211-3215.

Hepburn, H. R., & Radloff, S. E. (1988). *Honeybees of Africa*: Berlin, Germany: Springer.

Hoggart, C. J., Shriver, M. D., Kittles, R. A., Clayton, D. G., & McKeigue, P. M. (2004). Design and analysis of admixture mapping studies. *The American Journal of Human Genetics, 74*(5), 965-978.

Jensen, A. B., Palmer, K. A., Boomsma, J. J., & Pedersen, B. V. (2005). Varying degrees of *Apis mellifera ligustica* introgression in protected populations of the black honeybee, *Apis mellifera mellifera*, in northwest Europe. *Molecular Ecology, 14*(1), 93-106. doi: 10.1111/j.1365-294X.2004.02399.x

Jensen, A. B., & Pedersen, B. V. (2005). Honeybee conservation: a case story from Læsø island, Denmark *Beekeping and Conserving Biodiversity of Honeybees* (pp. 142-164): Northern Bee Books.

Kalinowski ST, Manlove KR, & ML, T. (2007). ONCOR A computer program for Genetic Stock Identification. *Available: Department of Ecology, Montana State University, Bozeman MT 59717. Accessed: http://www.montana.edu/kalinowski.*

Kosoy, R., Nassir, R., Tian, C., White, P. A., Butler, L. M., Silva, G., Kittles, R., Alarcon-Riquelme, M. E., Gregersen, P. K., & Belmont, J. W. (2009). Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Human mutation, 30*(1), 69-78.

Maul, V., & Hähnle, A. (1994). Morphometric studies with pure bred stock of *Apis mellifera carnica* Pollmann from Hessen. *Apidologie, 25*(2), 119-132.

Meixner, M. D., Costa, C., Kryger, P., Hatjina, F., Bouga, M., Ivanova, E., & Büchler, R. (2010). Conserving diversity and vitality for honey bee breeding. *Journal of Apicultural Research, 49*(1), 85-92.

Meixner, M. D., Leta, M. A., Koeniger, N., & Fuchs, S. (2011). The honey bees of Ethiopia represent a new subspecies of *Apis mellifera—Apis mellifera simensis* n. ssp. *Apidologie, 42*(3), 425-437.

Meixner, M. D., Pinto, M. A., Bouga, M., Kryger, P., Ivanova, E., & Fuchs, S. (2013). Standard methods for characterising subspecies and ecotypes of *Apis mellifera*. *Journal of Apicultural Research, 52*(4), 1-28.

Miguel, I., Iriondo, M., Garnery, L., Sheppard, W., & Estonba, A. (2007). Gene flow within the M evolutionary lineage of *Apis mellifera*: role of the Pyrenees, isolation by distance and post-glacial re-colonization routes in the western Europe. *Apidologie, 38*(2), 141-155.

Muñoz, I., Dall'Olio, R., Lodesani, M., & De la Rúa, P. (2009). Population genetic structure of coastal Croatian honeybees (*Apis mellifera carnica*). *Apidologie, 40*(6), 617-626.

Nedić, N., Francis, R. M., Stanisavljević, L., Pihler, I., Kezić, N., Bendixen, C., & Kryger, P. (2014). Detecting population admixture in honey bees of Serbia. *Journal of Apicultural Research, 53*(2), 303-313.

Nolte, A. W., & Tautz, D. (2010). Understanding the onset of hybrid speciation. *Trends in Genetics, 26*(2), 54-58.

Oleksa, A., Chybicki, I., Tofilski, A., & Burczyk, J. (2011). Nuclear and mitochondrial patterns of introgression into native Dark bees (*Apis mellifera mellifera*) in Poland. *Journal of Apicultural Research, 50*(2), 116-129.

Ozerov, M., Vasemägi, A., Wennevik, V., Diaz-Fernandez, R., Kent, M., Gilbey, J., Prusov, S., Niemelä, E., & Vähä, J.-P. (2013). Finding markers that make a difference: DNA pooling and SNP-arrays identify population informative markers for genetic stock identification. *PLOS ONE, 8*(12), e82434. doi: 10.1371/journal.pone.0082434

Pardo-Seco, J., Martinón-Torres, F., & Salas, A. (2014). Evaluating the accuracy of AIM panels at quantifying genome ancestry. *BMC genomics, 15*(1), 543.

Parra, E. J., Marcini, A., Akey, J., Martinson, J., Batzer, M. A., Cooper, R., Forrester, T., Allison, D. B., Deka, R., & Ferrell, R. E. (1998). Estimating African American admixture proportions by use of population-specific alleles. *The American Journal of Human Genetics, 63*(6), 1839-1851.

Paschou, P., Ziv, E., Burchard, E. G., Choudhry, S., Rodriguez-Cintron, W., Mahoney, M. W., & Drineas, P. (2007). PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS genetics, 3*(9), e160.

Pinto, M. A., Henriques, D., Chávez-Galarza, J., Kryger, P., Garnery, L., van der Zee, R., Dahle, B., Soland-Reckeweg, G., De la Rúa, P., & Dall'Olio, R. (2014). Genetic integrity of the Dark European honey bee (*Apis mellifera mellifera*) from protected populations: a genome-wide assessment using SNPs and mtDNA sequence data. *Journal of Apicultural Research, 53*(2), 269-278.

Pinto, M. A., Muñoz, I., Chávez-Galarza, J., & De la Rúa, P. (2012). The Atlantic side of the Iberian Peninsula: a hot-spot of novel African honey bee maternal diversity. *Apidologie, 43*(6), 663-673.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., & Daly, M. J. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics, 81*(3), 559-575.

Raymond, M., & Rousset, F. (2004). GENEPOP (version 3.4): Population genetics software for exact tests and ecumenicism. Laboratoire de Genetique et Environment, Montpellier, France.

Rhymer, J. M., & Simberloff, D. (1996). Extinction by hybridization and introgression. *Annual Review of Ecology and Systematics, 27*(1), 83-109.

Rortais, A., Arnold, G., Alburaki, M., Legout, H., & Garnery, L. (2011). Review of the DraI COI-COII test for the conservation of the black honeybee (*Apis mellifera mellifera*). *Conservation Genetics Resources, 3*(2), 383-391.

Rosenberg, N. A., Li, L. M., Ward, R., & Pritchard, J. K. (2003). Informativeness of genetic markers for inference of ancestry. *The American Journal of Human Genetics, 73*(6), 1402-1422.

Ruttner, F. (1988). *Biogeography and taxonomy of honeybees*: 1st ed. Berlin, Germany: Springer Verlag.

Sambrook, J., & Russell, D. W. (1989). Molecular cloning: a laboratory manual. second. *Cold pring Harbor Laboratory Press, New York*.

Sheppard, W. S., & Meixner, M. D. (2003). *Apis mellifera pomonella*, a new honey bee subspecies from Central Asia. *Apidologie, 34*(4), 367-375.

Soland-Reckeweg, G., Heckel, G., Neumann, P., Fluri, P., & Excoffier, L. (2009). Gene flow in admixed populations and implications for the conservation of the Western honeybee, *Apis mellifera*. *Journal of Insect Conservation, 13*(3), 317.

Storer, C. G., Pascal, C. E., Roberts, S. B., Templin, W. D., Seeb, L. W., & Seeb, J. E. (2012). Rank and order: valuating the performance of SNPs for individual assignment in a non-model organism. *PLOS ONE, 7*(11), e49018. doi: 10.1371/journal.pone.0049018

Strange, J. P., Garnery, L., & Sheppard, W. S. (2008). Morphological and molecular characterization of the Landes honey bee (*Apis mellifera L.*) ecotype for genetic conservation. *Journal of Insect Conservation, 12*(5), 527-537.

Uzunov, A., Meixner, M. D., Kiprijanovska, H., Andonov, S., Gregorc, A., Ivanova, E., Bouga, M., Dobi, P., Büchler, R., & Francis, R. (2014). Genetic structure of *Apis mellifera* macedonica in the Balkan Peninsula based on microsatellite DNA polymorphism. *Journal of Apicultural Research, 53*(2), 288-295.

vanEngelsdorp, D., & Meixner, M. D. (2010). A historical review of managed honey bee populations in Europe and the United States and the factors that may affect them. *Journal of invertebrate pathology, 103*, S80-S95.

Wallberg, A., Han, F., Wellhagen, G., Dahle, B., Kawata, M., Haddad, N., Simões, Z. L. P., Allsopp, M. H., Kandemir, I., & De la Rúa, P. (2014). A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee *Apis mellifera*. *Nature genetics, 46*(10), 1081-1088.

Weinstock, G. M., Robinson, G. E., Gibbs, R. A., Worley, K. C., Evans, J. D., Maleszka, R., Robertson, H. M., Weaver, D. B., Beye, M., & Bork, P. (2006). Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature, 443*(7114), 931-949.

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *evolution, 38*(6), 1358-1370.

Whitfield, C. W., Behura, S. K., Berlocher, S. H., Clark, A. G., Johnston, J. S., Sheppard, W. S., Smith, D. R., Suarez, A. V., Weaver, D., & Tsutsui, N. D. (2006). Thrice out of Africa: ancient and recent expansions of the honey bee, *Apis mellifera*. *Science, 314*(5799), 642-645.

Wilkinson, S., Archibald, A. L., Haley, C. S., Megens, H.-J., Crooijmans, R. P., Groenen, M. A., Wiener, P., & Ogden, R. (2012). Development of a genetic tool for product regulation in the diverse British pig breed market. *BMC genomics, 13*(1), 580.

Wilkinson, S., Wiener, P., Archibald, A. L., Law, A., Schnabel, R. D., McKay, S. D., Taylor, J. F., & Ogden, R. (2011). Evaluation of approaches for identifying population informative markers from high density SNP chips. *BMC Genetics, 12*(1), 45. doi: 10.1186/1471-2156-12-45

Wright, S. (1978). *Evolution and the genetics of populations, variability within and among populations*.

Chapter IV.

High sample throughput genotyping for estimating C-lineage introgression in the Dark honey bee: an accurate and cost-effective SNP-based tool

The paper was submitted to the Journal *Scientific Report* s:

Dora Henriques, Keith A. Browne, Mark W. Barnett, Melanie Parejo, Per Kryger, Tom C. Freeman, Irene Muñoz, Lionel Garnery, Fiona Highet, J. Spencer Jonhston, Grace P. McCormack, M. Alice Pinto

Abstract

The natural distribution of the honey bee (*Apis mellifera* L.) has been changed by humans in recent decades to such an extent that the formerly widest-spread European subspecies, *Apis mellifera mellifera*, is threatened by extinction through introgression from highly divergent commercial strains in large tracts of its range. Conservation efforts for *A. m. mellifera* are underway in multiple European countries requiring reliable and cost-efficient molecular tools to identify purebred colonies. Here, we developed four ancestry-informative SNP assays for high sample throughput genotyping using the iPLEX Mass Array system. Our customized assays were tested on DNA from individual and pooled, haploid and diploid honey bee samples extracted from different tissues using a diverse range of protocols. The assays had a high genotyping success rate and yielded accurate genotypes. Performance assessed against whole-genome data showed that individual assays behaved well, although the most accurate introgression estimates were obtained for the four assays combined (117 SNPs). The best compromise between accuracy and genotyping costs was achieved when combining two assays (62 SNPs). We provide a ready-to-use cost-effective tool for accurate molecular identification and estimation of introgression levels to more effectively monitor and manage *A. m. mellifera* conservatories.

Keywords: *Apis mellifera mellifera*, SNP assays, introgression, conservation

Introduction

Pollination by the honey bee (*Apis mellifera* L.) is a blended ecosystem service of managed and unmanaged (feral or wild) colonies that is under threat from human-mediated environmental changes including climate change, habitat loss, habitat fragmentation, pesticides, and introduced parasites and pathogens (Potts *et al.* 2010; vanEngelsdorp & Meixner, 2010). There is growing evidence that management of locally adapted genetic diversity in honey bee subspecies and ecotypes is key to the long-term sustainability of this service (Büchler *et al.*, 2014; Meixner *et al.*, 2014; Meixner *et al.*, 2015). Accordingly, actions towards preserving the large stores of genetic diversity held by the 31 honey bee subspecies (Chen *et al.*, 2016; Engel, 1999; Meixner *et al.*, 2011; Sheppard & Meixner, 2003) are expected to counteract the trend of global colony losses.

Of the 31 subspecies that have been identified in the natural distributional range of *A. mellifera* in Africa, Middle East, Western Asia, and Europe (Chen *et al.*, 2016; Ruttner, 1988 ; Sheppard & Meixner, 2003) there are 10 European subspecies grouped into two evolutionary lineages (Ruttner, 1988): the Western and Northern European (lineage M) and the Southeastern European (lineage C). Lineage M includes only two subspecies: the Dark honey bee, *Apis mellifera mellifera*, and the Iberian honey bee *Apis mellifera iberiensis*. Yet, these two subspecies cover the largest territory in Europe with *A. m. iberiensis* occupying the Iberian Peninsula and *A. m. mellifera* ranging from France in the south to Scandinavia in the north, and from Ireland and the UK in the west to the Ural Mountains in the east (Ruttner, 1988). Lineage C occurs in a smaller geographical area composed of the Apennine and Balkan peninsulas and includes the most widely kept honey bee subspecies: the Italian *Apis mellifera ligustica* and the Carniolan *Apis mellifera carnica*. In spite of its wide distribution, *A. m. mellifera* is the subspecies most under threat as it is considered extinct in many parts of Europe not only because of the human-mediated environmental changes but more insidiously through replacement by and introgression from non-indigenous subspecies, particularly *A. m. ligustica* and *A. m. carnica* (Jensen *et al.*, 2005; Pinto *et al.*, 2014; Soland-Reckeweg *et al.*, 2008).

It has been argued that, unlike with other domesticated stock organisms, management and selective breeding in honey bees increase genetic diversity through introgression (Harpur *et al.*, 2014). However, this form of admixture reduces the frequency of locally adapted gene complexes, leading to an increased likelihood of reduced survival rates of colonies (De la Rúa *et al.*, 2013).

How to protect locally adapted gene complexes that are more suited to local environments is a growing problem, as the increased breeding and movement of C-lineage honey bees promotes sympatry and gene flow between *A. m. mellifera* and imported commercial breeds. Efforts to assist conservation of *A. m. mellifera* are gathering momentum in multiple European countries (www.sicamm.org) and with the knowledge that reduced adapted genetic diversity threatens both managed and unmanaged populations, the interests of commercial beekeeping and honey bee conservationists should be aligning, particularly in *A. m. mellifera* indigenous areas.

An important first step in protecting *A. m. mellifera* populations in official or unofficial conservatories is to give the stakeholders an accurate and cost-efficient tool to test for C-lineage introgression. Microsatellites have been extensively used to examine C-lineage introgression in *A. m. mellifera* (Jensen *et al.*, 2005; Meixner *et al.*, 2013; Soland-Reckeweg *et al.*, 2008). Yet, the numerous advantages of SNPs over microsatellites promise to make them the tool of choice for population monitoring and conservation purposes. In addition to being more abundant and widespread in the genome (Weinstock *et al.*, 2006), SNPs display lower genotyping error, have higher quality data, are more amenable to automated analysis and data interpretation, and can be easily transferred between laboratories (Vignal *et al.*, 2002). Moreover, SNPs proved to be more powerful than microsatellites at estimating C-lineage introgression in *A. m. mellifera* (Muñoz *et al.*, 2017). These properties make SNPs a powerful tool for testing the breeding stock in *A. m. mellifera* conservatories and SNP data can be readily incorporated in shared genetic databases, facilitating implementation of a conservation strategy at the European scale.

Whilst SNP analysis on whole genome (WG) sequence data may be required in studies concerned with fine-scale relatedness, such deep sequencing is disproportionate when determining introgression levels for the discrimination of *A. m. mellifera* breeding stocks. Also, while costs have dropped dramatically, WG sequencing is still unaffordable for most stakeholders committed to the long-term sustainability and conservation of honey bees. Costs are accrued as WG analysis requires considerable computing storage and processing power and trained bioinformatics personnel. However, encouragingly, recent studies showed that reduced panels of highly informative SNPs can accurately identify honey bee stocks (Chapman *et al.*, 2017; Chapman *et al.*, 2015; Muñoz *et al.*, 2015; Parejo *et al.*, 2016). Genotyping using reduced SNP panels considerably decreases laboratory processing costs. Furthermore, analysis of the generated genotypes requires low computational power and conventional bioinformatics skills.

Muñoz *et al.* (2015) developed reduced SNP panels for genetic identification and introgression analysis in *A. m. mellifera.* The authors used a combination of metrics to rank by information content over 1183 SNPs that had been genotyped in *A. m. mellifera*, *A. m. ligustica* and *A. m. carnica* using the 1536-plex GoldenGate® Assay of Illumina (Pinto *et al.*, 2014). The top-ranked SNPs were combined into five nested panels whose sizes (48, 96, 144, 192, 384 SNPs each) fitted the plexes of the now discontinued GoldenGate® Assays formerly genotyped with the VeraCode® technology. Here, we built from the 144-SNP panel to propose four customized assays tailored for high sample throughput genotyping using the iPLEX MassARRAY system. By providing a ready-to-use molecular tool for accurately, rapidly, and cost-effectively genotyping large sample sizes of *A. m. mellifera*, we hope to bring affordable C-lineage introgression detection to stakeholders in the fight to safeguard remaining reservoirs of unique combinations of genes and adaptations in *A. m. mellifera* and to expand its reduced current distribution.

Results

Assay design, quality control and genotyping accuracy

Of the 144 highly-informative SNPs selected by Muñoz *et al.* (2015), the Assay Design software was able to multiplex 127 into four assays (identified by letter M), each containing a variable number of SNPs ranging from 38 in M1 to 24 in M4 (Table Sup IV-1). A total of 573 samples (Figure IV-1) were genotyped for the four customized assays using the iPLEX MassARRAY system. Of the 573 samples, only seven displayed a SNP call failure rate >30% and these were excluded from further analysis (Table Sup IV-2). Of the 566 remaining samples, 551 displayed a low percentage (<10%) of missing data indicating a high genotyping success rate (96%).

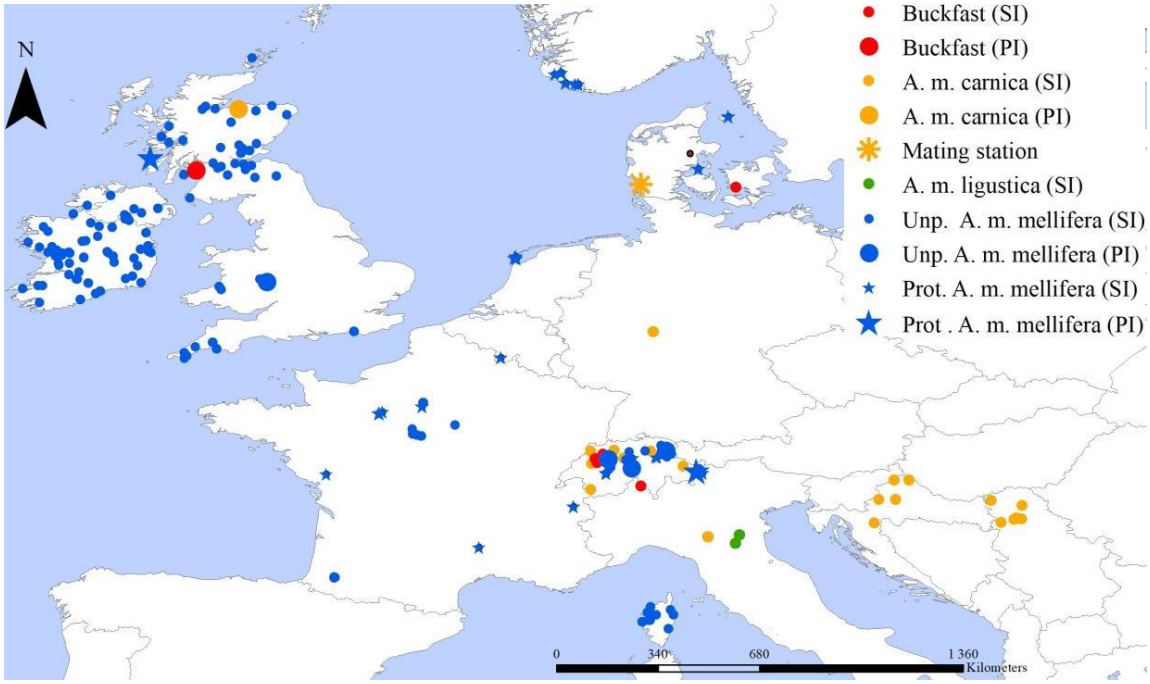


Figure IV-1 - Location of the colonies sampled across the *A. m. mellifera* and C-lineage ranges. Samples of *A. m. mellifera* were collected in protected (Prot) and unprotected apiaries (Unp). Colonies were genotyped for the four SNP assays in the MassARRAY® MALDI-TOF platform from single individuals (SI) or pools of individuals (PI).

The quality control and assessment of the genotyping accuracy of the 127 SNPs (Table Sup IV- 3) led to identification of 10 problematic SNPs, of which seven were typed in <80% of the individuals, three were called heterozygous for >10% of the haploid individuals (Tables Sup IV-1 and Sup IV-3), and three exhibited inconsistent calls among the three genotyping technologies in >5% of the individuals (Figure IV-2). The latter SNPs were also identified as having high rates of missing data or heterozygosity (Table Sup IV-3). Once the 10 SNPs were removed from the datasets, the rates of missing data of the remaining 117 SNPs were low with 113 having <10% and four varying between 10.4% and 15.5% (Table Sup IV 1). The genotypes generated for the 117 high-quality SNPs in the MassARRAY platform were highly concordant with those of the Illumina's platforms (99.9% for the BeadArray and 99.6% for the HiSeq 2500). Following the quality control step, 339 of the 573 genotyped samples had no missing data and the highest rate of missing data was 29% but only in two samples (Table Sup IV-2).

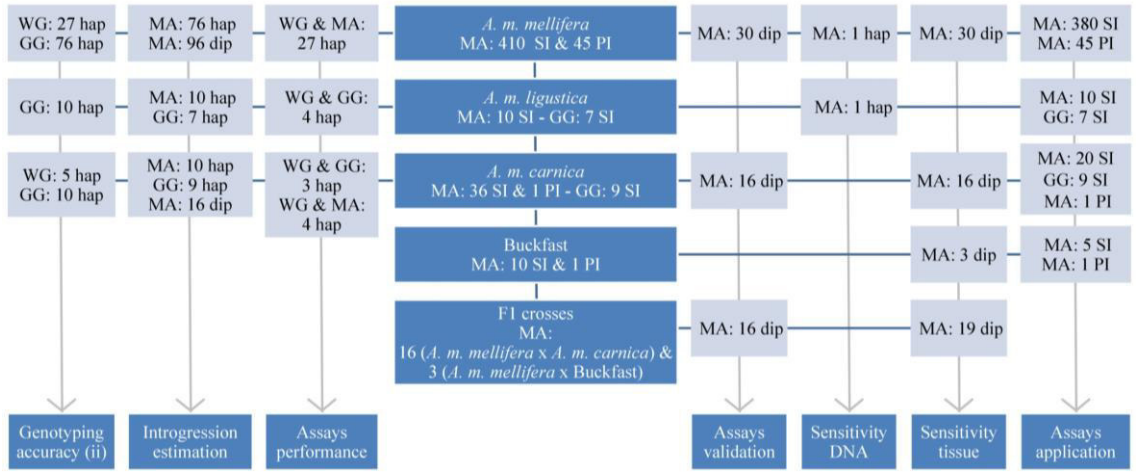


Figure IV-2 - Datasets of quality-proved samples used in the SNP assays' testing and application. Samples were represented by a single individual (SI) or a pool of individuals (PI). The individuals were haploid drones (hap) or diploid workers (dip). Genotypes were generated from the four assays in the MassARRAY® MALDI-TOF platform (MA), from the GoldenGate® Assay in the Illumina's BeadArray platform (GG), and from whole genome (WG) sequences in the Illumina's HiSeq 2500 platform.

The final multiplexes contained M1=34, M2=32, M3=28, and M4=23 SNPs distributed across the 16 honey bee linkage groups, LG (Figure IV-3; Tables Sup IV-1 and Sup IV-4). LG 2 harboured the highest number of SNPs (13) while LG 3 had the lowest (2). The number of LGs covered by the assays varied between 12 (M4) and 14 (M1). Most SNPs (90 of 117) are located in non-coding regions, including intergenic (50 SNPs), intronic (30 SNPs), and UTRs (10 SNPs). Of the 27 coding SNPs, only two (1384-est6107 and 661-AMB-00398036) are non-synonymous (Table Sup IV-1).

Assessing introgression estimation

Estimates of C-lineage introgression proportions (*Q*-values) into *A. m. mellifera* were produced by STRUCTURE and ADMIXTURE and using datasets of varying ploidies (Figure IV-2). As expected, the two software packages estimated virtually the same *Q*-values (*P*-value=0.89, Mann-Whitney test; Table Sup IV-5). Similar *Q*-values were also inferred from haploid, diploid, and combined haplodiploid datasets (*P*-value>0.87, Mann-Whitney test; Table Sup IV-6).

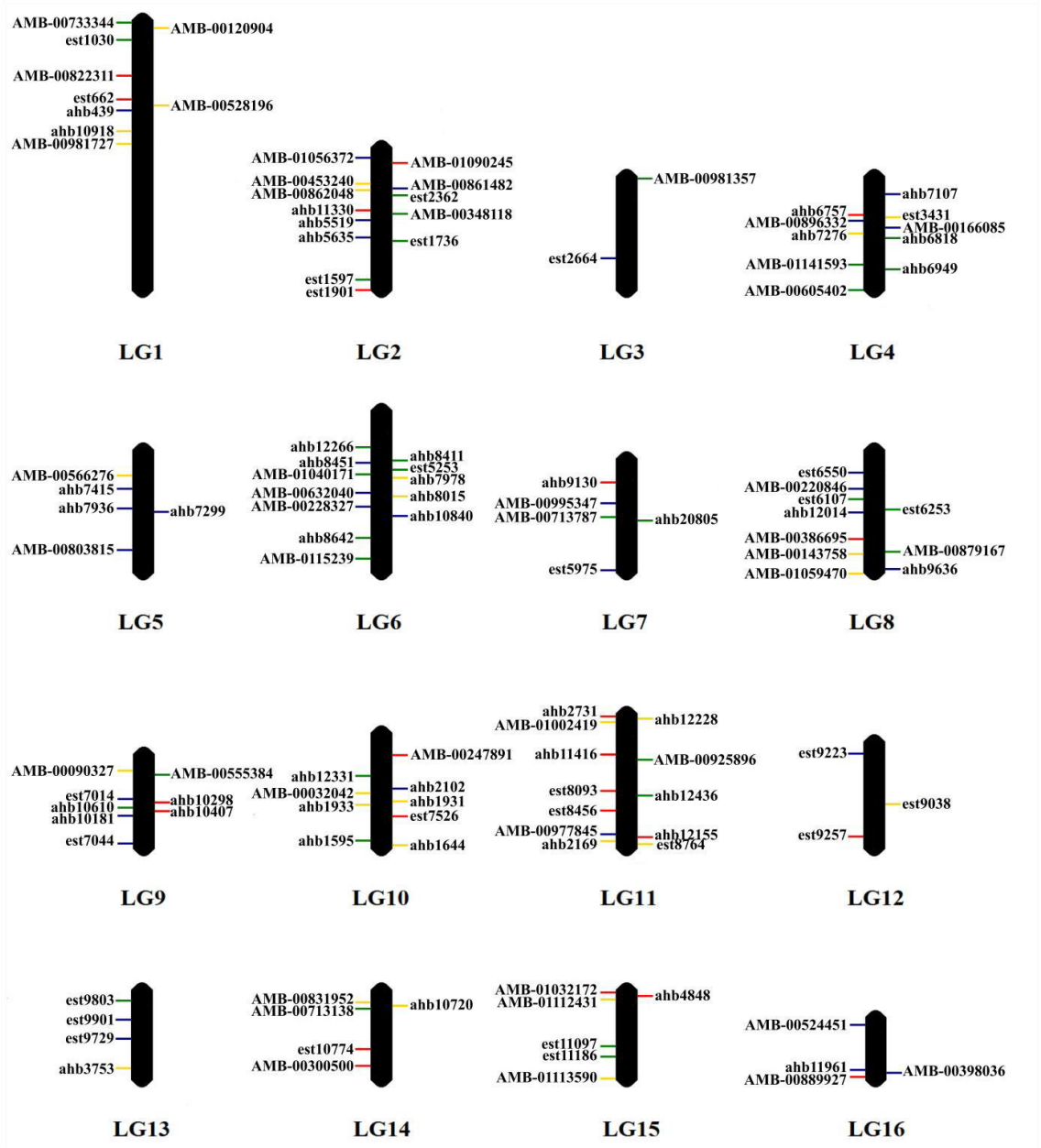


Figure IV-3 – Genomic positions of the 117 quality-proved SNPs. The 117 SNPs were multiplexed in four assays, named M1 (blue), M2 (green), M3 (yellow), and M4 (red).

Assessing performance of the SNP assays

The performance of the four assays was assessed by comparing their *Q*-values (inferred from single or combined assays) with those inferred from the genome-wide SNPs, which provides the best estimate of the admixture proportions (Table Sup IV-7). The four assays exhibited a good individual performance with a mean accuracy >94% and *Q*-values highly correlated ($0.980 \leq r \leq 0.983$) with those inferred from the WG dataset (Table IV-1). The largest plex assay M1 (34 SNPs) and the smallest M4 (23 SNPs) showed the best and the worst behaviour, respectively,

as indicated by most statistics (Table IV-1)). The best performance was achieved when the four assays were used together (r=0.996; mean accuracy=97.84%; absolute precision error=0.033), although the combination of the two (M1+M3) and the three assays (M1+M2+M3) with the highest individual correlations produced equally interesting statistics with mean accuracies >96.9%, absolute precision error <0.04, and with over 28 individuals (out of 32) with absolute accuracy error <0.05. Performance was also assessed by counting purebred *A. m. mellifera* individuals misclassified as admixed (Q-values>0.05) and *vice versa* (Table IV-1). Except for M4, single assays and their combinations repeatedly misclassified two or three (always identified amongst individuals M23, M24, M25, and M26; Table Sup IV-7) purebred as admixed from 11 *A. m. mellifera* individuals identified by genome-wide SNPs. The degree of *A. m. mellifera* misclassification was lower for the class "admixed identified as purebred" with M3, and its combination with one (M1), two (M1+M2) or three assays (M1+M2+M4) correctly identifying all 16 admixed individuals (0.05<Q-value<0.95).

Table IV-1 - Statistics for the performance of the four SNP assays used singly or combined. Calculations were made via comparisons between Q-values inferred from the SNP assays and the genome-wide 2.399 million SNPs. (i) Pearson's correlation coefficient (r); (ii) similarity score obtained by CLUMPAK; (iii) mean and (iv) maximum absolute accuracy errors; (v) number of individuals (out of 38) with absolute accuracy error <0.05; (vi) mean accuracy estimated via percentage of absolute error; (vii) absolute precision error; (viii) number of purebred *A. m. mellifera* individuals misclassified as admixed; (ix) number of admixed individuals misclassified as purebred.

SNP Assay	**# of SNPs**	**(i)**	**(ii)**	**(iii)**	**(iv)**	**(v)**	**(vi)**	**(vii)**	**(viii)**	**(ix)**
M1	34	0.983	0.929	0.046	0.211	26	95.42	0.061	2	1
M2	32	0.981	0.919	0.051	0.239	24	94.86	0.068	3	2
M3	28	0.982	0.926	0.047	0.314	23	95.27	0.066	3	0
M4	23	0.980	0.911	0.050	0.283	23	95.00	0.067	0	2
M1+M3	62	0.993	0.956	0.029	0.172	31	97.09	0.042	2	0
M1+M2+M3	94	0.994	0.957	0.031	0.137	28	96.94	0.040	3	0
M1+M2+M3+M4	117	0.996	0.964	0.022	0.114	32	97.84	0.033	2	0

Validating the SNP assays

The assays were validated using an independent set of 62 individuals, including 30 *A. m. mellifera*, 16 *A. m. carnica*, and 16 F1 hybrids. On average, *Q*-values inferred from the observed genotypes called using the four (individual or combined) assays were similar to the expected *Q*-values (Figure IV-4). Despite good overall performance of the assays, a few purebred *A. m. mellifera* and *A. m. carnica* were misclassified as admixed (estimated *Q*-values deviated from thresholds of <0.05 for *A. m. mellifera* and >0.95 for *A. m. carnica*) when the *Q*-values were inferred from observed genotypes called using individual assays (Figure IV-4, Table Sup IV-8). However, when variable combinations of the four assays were employed, the estimated *Q*-values matched the expectations with all *A. m. mellifera* and *A. m. carnica* correctly classified as purebred and the F1 hybrids varying between 0.52±0.04 (mean±SD), for M1+M3, and 0.56±0.03, for the four assays combined.

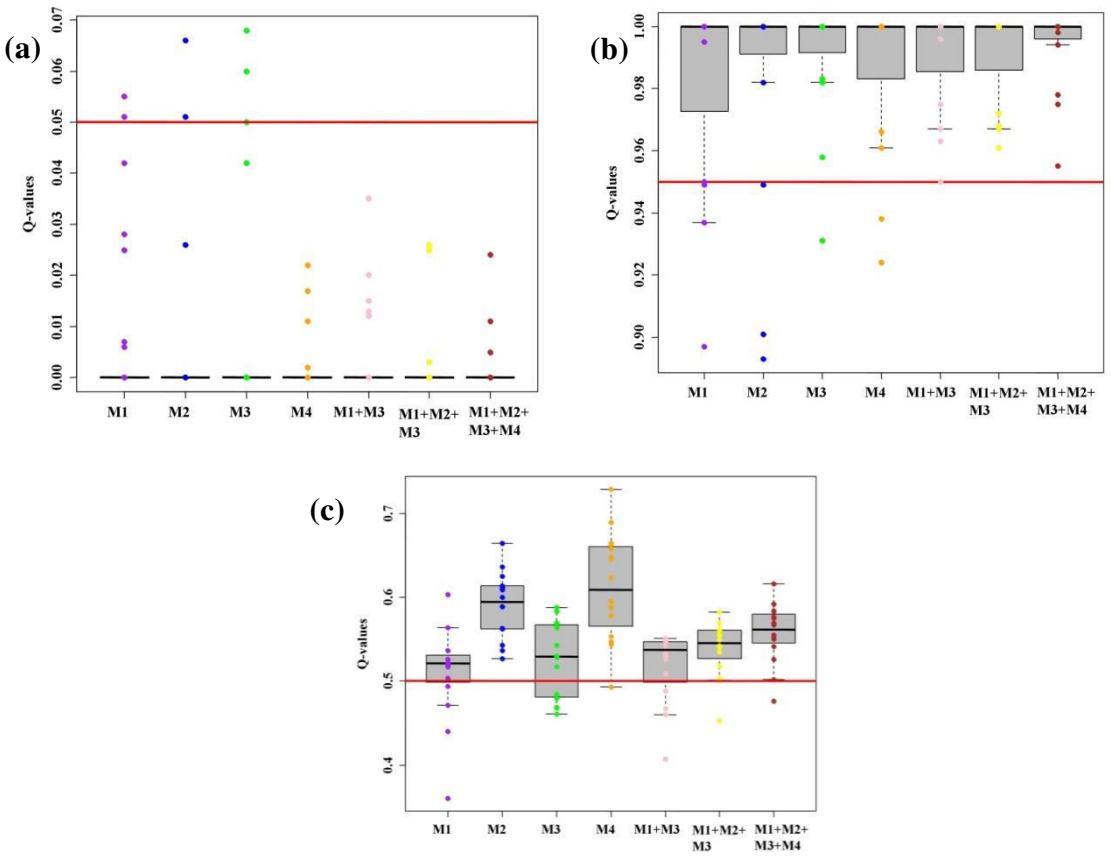


Figure IV-4 - Validating the four SNP assays. Boxplots showing the variation of the Q-values inferred from the observed genotypes for the four SNP assays. The boxes denote the first and third quartiles. The horizontal red lines mark the expected *Q*-values for purebred *A. m. mellifera and A. m. carnica* set at <0.05 and >0.95, respectively, and for the F1 hybrid samples set at 0.5. Boxplots for the **(a)** 30 *A. m. mellifera* samples, **(b)** 16 *A. m. carnica* samples, and **(c)** 16 F1 hybrid samples.

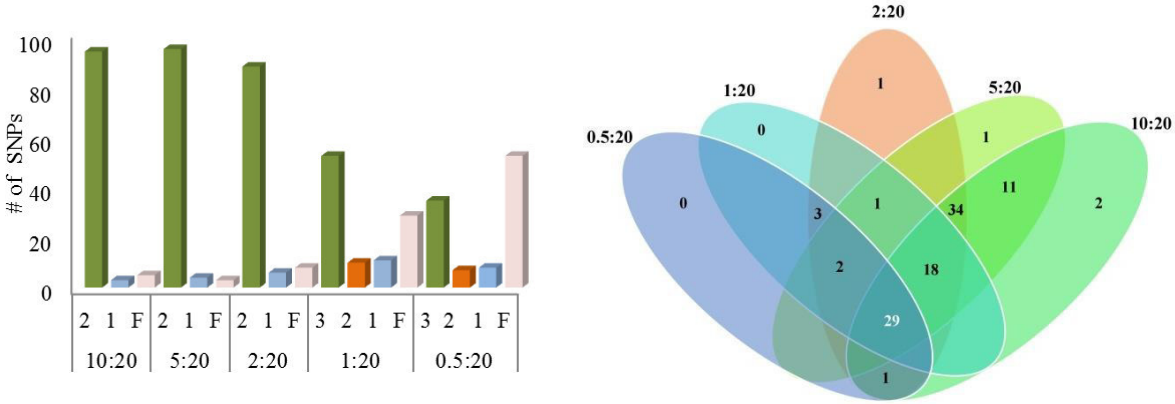


Figure IV-5 – Sensitivity of the MassARRAY genotyping system assessed in pooled DNA. (a) Number of correctly called and failed (F) SNP loci across dilution ratios (10:20, 5:20, 2:20, 1:20, 0.5:20) and replicates (1, 2, 3). (b) Venn diagram of the number of SNP loci with 100% successful SNP calls at each dilution ratio. The central overlap shows the 29 SNPs that resulted in 100% success for all dilution ratios.

Assessing sensitivity of the MassARRAY system in pooled DNA

The sensitivity of the MassARRAY genotyping system was assessed in pools combining the DNA of two haploid individuals (one *A. m. mellifera* and one *A. m. ligustica*) at five dilution ratios. Of the 117 SNP loci, 14 were uninformative for two different reasons; while five SNP loci were monomorphic in the two individuals, nine were bi-allelic but only one allele (either the *A. m. mellifera* or the *A. m. ligustica* allele) was called across the five dilution ratios. The sensitivity of the MassARRAY platform in detecting the alleles of *A. m. ligustica* was assessed in the remaining 103 loci. As expected, the power of the genotyping system in detecting *A. m. ligustica* alleles decreased as the dilution ratios increased (Figure IV-5a). At the ratios of 10:20 and 5:20, the *A. m. ligustica* alleles were detected for 93 SNP loci, a number that decreased substantially for the dilution 0.5:20, with alleles detected only for 32 SNPs. The MassARRAY platform was able to detect the *A. m. ligustica* alleles in every dilution ratio and replicate in only 29 SNPs (Figure IV-5b; Table Sup IV-1).

The average *Q*-values inferred from the observed genotypes for each dilution using the four assays (117 SNPs), the two best assays M1+M3 (62 SNPs), and cross-detected SNPs (29 SNPs) are shown in Figure IV-6. Despite the considerably lower number of SNP loci genotyped with M1+M3, the *Q*-values inferred from them were highly correlated (r=0.99) with those inferred from the four assays (Table Sup IV-9). As expected, the *Q*-values inferred from the genotypes called with the 29 SNPs were always 0.50, as these were all heterozygous. The *Q*-values inferred from the

observed genotypes obtained for the DNA pools of 10:20 (0.471≤ *Q*-value ≤0.475) and 5:20 (0.418≤ *Q*-value ≤0.455) were similar to those inferred from the expected genotypes (0.425≤ *Q*-value ≤0.444) using either the four assays or M1+M3 (Table Sup IV-9). As the dilution ratios decreased so did the *Q*-values reaching zero at 0.5:20, indicating that at such low concentration of the *A. m. ligustica* DNA the MassARRAY platform was unable to detect C-derived alleles.

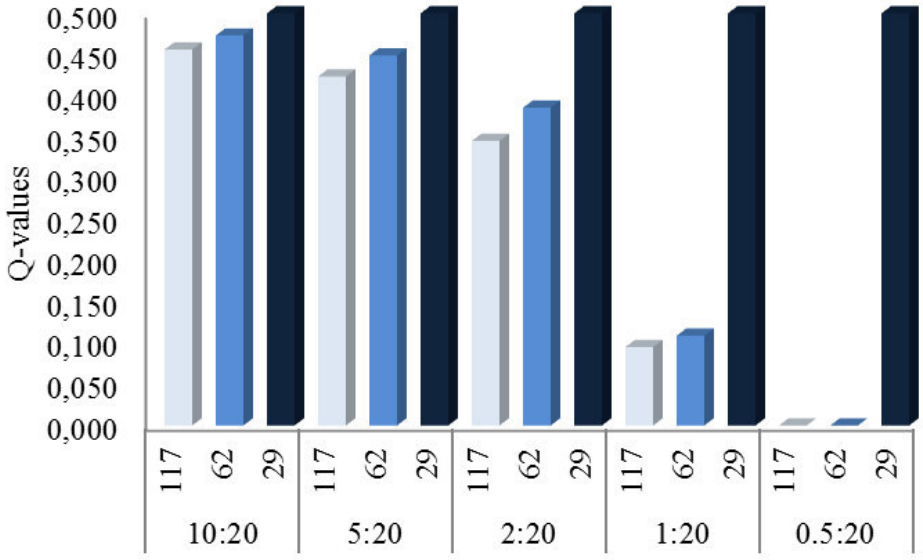


Figure IV-6 - Average *Q*-values for different DNA pools. *Q*-values were inferred for DNA pools (representing dilution ratios of 10:20, 5:20, 2:20, 1:20, 0.5:20) by the four SNP assays (117 SNPs), the two best assays M1+M3 (62 SNPs) and the 29 SNPs that were identified in all dilution ratios.

Assessing sensitivity of the MassARRAY system in pooled tissue

The sensitivity of the MassARRAY genotyping system was further assessed in 22 tissue pools. A total of 2,574 genotypes (117 SNP loci x 22 pools) were called by the MassARRAY platform. Of these, 1,977 (77%) were accurate, as determined by comparing the calls obtained for single workers with those obtained for the pools. The most common sources of mismatch were "the most frequent allele" (279 genotypes) and "higher DNA concentration" (86 genotypes; Table IV-2 and Table IV-3). For example, in the pool combining one F1 hybrid with seven *A. m. mellifera*, the genotype calls for locus ahb12014 were AG for the former, AA for the latter and AA for the pool. In another example, provided by the pool combining one F1 hybrid (DNA concentration=288.6 ng/µl; Table Sup IV-10) with one *A. m. mellifera* (DNA concentration=60.2 ng/µl; Table Sup IV-10), the genotype calls for locus ahb10407 were GG for the former, AA for the latter and GG for the pool. In this case the genotype of the pool was determined by the individual that had a DNA concentration five-fold higher in the individual extraction.

Table IV-2 - Information on SNP calling obtained from the 22 tissue pools.

SNP calling	# of genotypes
Sources of allele miscalling	
Different alleles	5
Higher DNA concentration	86
Higher DNA concentration & the most frequent allele	81
The most frequent allele	279
The least frequent allele	42
Missing data	104
Accurate calls	1,977
Total	2,574

The 117 SNP loci were accurately called in at least seven pools (Table Sup IV-11). This number decreased steadily as the pools increased being only 14 in the 22 pools, of which only three overlapped with the 29 SNPs identified in the pooled-DNA experiment. Nonetheless, the average rate of accurately called SNPs per pool was high (77%, 90 SNPs) and varied between 83% (97 SNPs), for the pools of two workers, and 50% (58 SNPs), for the pools of eight workers (Table IV-3; Table Sup IV-12).

Table IV-3 - Mean number of SNP loci accurately called and miscalled for the different combination of tissue pools. The sources of miscalling were (i) different alleles, (ii) higher DNA concentration, (iii) higher DNA concentration and the most frequent allele, (iv) the most frequent allele, and (vi) the least frequent allele. Mel - *A. m. mellifera*; Hyb – F1 hybrid; Car – *A. m. carnica*; Buc – Buckfast.

Tissue pools (# of replicates)	Accurate SNPs	Miscalled SNPs				
		i	ii	iii	iv	v
1 Mel + 1 Hyb (3)	97.0	1.0	4.7	4.7	5.3	0.3
2 Mel + 1 Hyb (2)	81.0	0.0	3.5	5.0	21.0	0.5
3 Mel + 1 Hyb (2)	83.5	0.0	5.5	5.0	17.5	0.0
7 Mel + 1 Hyb (2)	58.0	0.0	4.5	5.0	47.0	0.5
1 Mel + 1 Car (3)	85.0	0.3	4.3	4.0	5.3	11.0
1 Car + 1 Hyb (3)	101.7	0.3	3.7	2.0	4.0	1.3
2 Car + 1 Hyb (2)	93.5	0.0	3.0	2.5	13.0	0.0
3 Car + 1 Hyb (2)	93.0	0.0	3.0	5.0	12.0	0.0
1 Buc +1 Hyb (3)	102.7	0.0	3.0	1.3	4.7	0.7

The *Q*-values estimated for the 22 pools from the expected and observed genotypes called using the four assays, the combination of M1+M3, and the 29 SNPs identified in the pooled-DNA experiment are shown in Table Sup IV-13. The *Q*-values inferred from the expected genotypes varied between 0.44 and 0.51 for the pools combining a variable number of *A. m. mellifera* with

one F1 hybrid. The range increased to 0.59≤ *Q*-values ≤0.66 when *A. m. mellifera* was replaced by either *A. m. carnica* or Buckfast. Similar *Q*-values were inferred from the observed genotypes using the four assays (r=0.80), M1+M3 (r=0.80), and the 29 SNPs (r=0.74). Despite the variation around the *Q*-values (Table Sup IV-13), the MassARRAY platform was able to detect low frequency alleles, either of M-lineage (pools containing *A. m. mellifera)* or C-lineage ancestry (pools containing *A. m. carnica* or Buckfast), across all tissue pools.

Applying the SNP assays

The four assays were applied to 478 colonies of various ancestries, represented by single (431 colonies) or pooled individuals (47 colonies), collected in 13 European countries (Figure IV-1). The average *Q*-values estimated for *A. m. mellifera* colonies, represented by a single individual, indicated that introgression varies throughout Europe, ranging from 0 in Norway to 0.447±0.265 in Wales (Table Sup IV-14). The least introgressed *A. m. mellifera* colonies were from conservatories of Norway (0±0.000), Scotland (0.006±0.011), Netherlands (0.046±0.141) and Belgium (0.059±0.046), although unprotected populations of Ireland were also very pure (0.021±0.022). Populations of Denmark, France and Switzerland exhibited greater *Q*-values (0.148≤*Q*-value≤0.280) in both protected and unprotected populations (Figure IV-7a, Table Sup IV-14). Admixture proportions estimated for *A. m. ligustica* and *A. m. carnica* sampled from native and introduced ranges showed that they are very pure (0.972≤*Q*-value≤1.000), excepting for some Swiss colonies (0.750±0.296). The commercial breed Buckfast was mostly of C-derived ancestry (0.806±0.055).

The genotype data were further examined by network analysis. The correlation network graph shown in Figure IV-7b consisted of 5,522 edges and 418 nodes (samples). Samples with similar allele profiles clustered together. In total, three clusters were identified with cluster 1 containing 342 nodes (highest similarity to M-lineage), cluster 2 containing 58 nodes (highest similarity to C-lineage) and cluster 3 containing 18 nodes (highest rates of introgression). All samples from Norway, Ireland, Netherlands and Belgium were in cluster 1 whilst all samples from Italy, Croatia and Serbia were in cluster 2. Of 70 samples from Scotland, 61 samples were in cluster 1, 6 in cluster 2 and only 2 in cluster 3; a similar distribution was seen for samples from France and Switzerland. Samples from England, Denmark and Wales were also predominantly found in cluster 1.

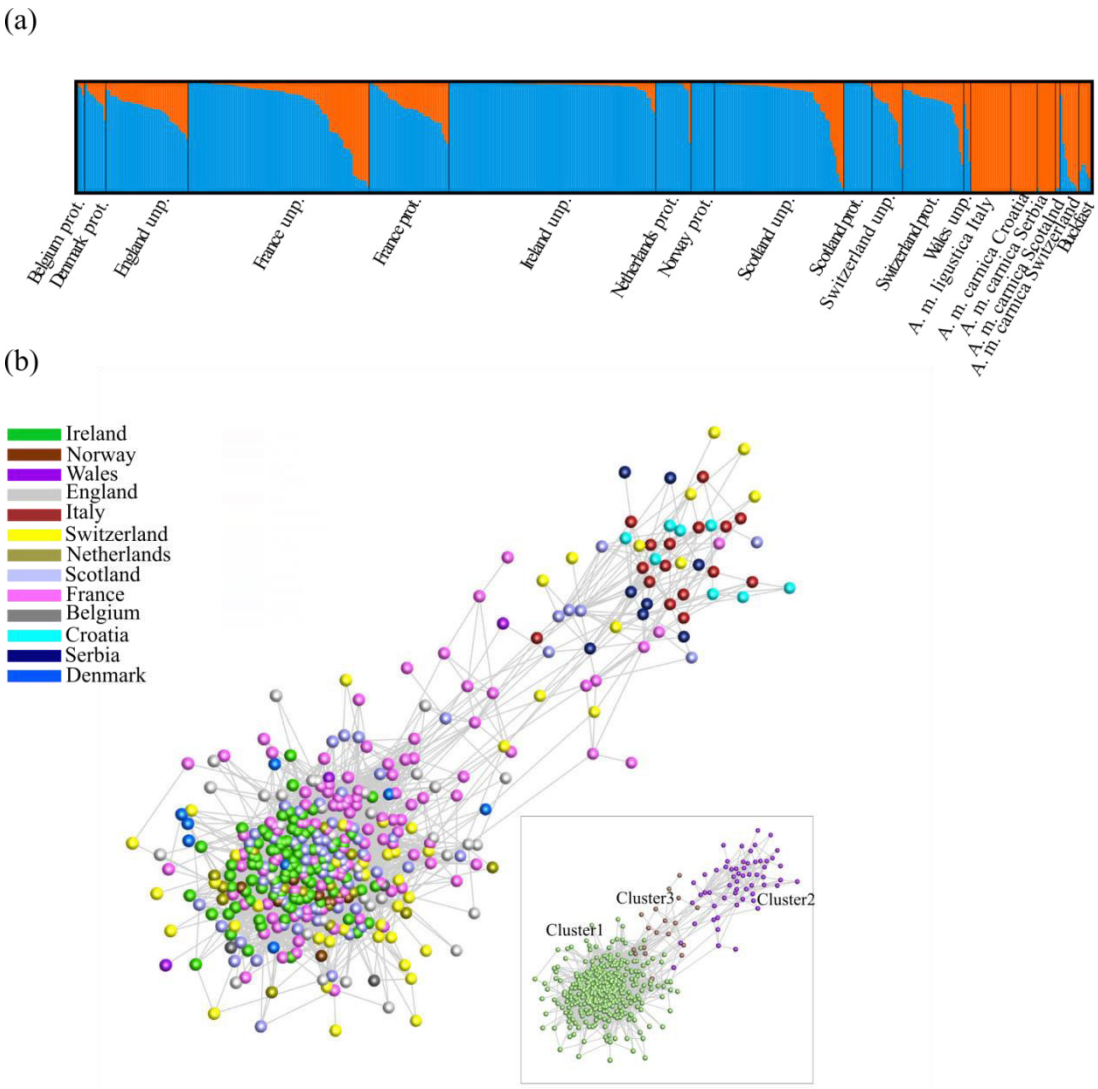


Figure IV-7 - Structure reconstructed by ADMIXTURE and Graphia Professional software packages for honey bees of diverse ancestry collected across Europe. Most depicted samples (415) were genotyped in the MassARRAY platform using the four assays (117 SNPs). Nine samples of *A. m. carnica* and seven *A. m. ligustica*, previously genotyped for the 117 SNP loci using the GoldenGate Assay in the BeadArray platform, were added to the structure analysis for a better representation of C-lineage diversity. Each sample corresponds to a single colony. Samples collected in the *A. m. mellifera* range are from protected (prot) and unprotected (unp) apiaries. **(a)** ADMIXTURE plot showing the genome partitioning into two clusters (K=2) for each individual, represented by a vertical bar. Blue represents the *A. m. mellifera* cluster and orange the C-lineage cluster. The black lines separate individuals from different countries and studied groups. (**b)** Correlation network where nodes (honey bee samples) are connected with edges when r>0.27. A total of 418 samples out of 431 formed connections in the graph. Samples coloured according to country of origin. Inset shows correlation network clustered using the Markov Cluster (MCL) algorithm at an inflation value of 1.2.

The admixture patterns were also examined in pooled individuals representing an independent set of 47 colony samples from Switzerland and the UK (Table Sup IV-15). The average *Q*-values estimated for the Swiss samples of *A. m. mellifera* varied between 0.145±0.074 (protected) and 0.118±0.042 (unprotected), which were lower than those inferred from a single individual (Table Sup IV-14). However, these estimates are not directly comparable as the pooled- and single-individual samples were from different apiaries. More comparable results were obtained for four colonies of variable ancestry from the UK that were simultaneously represented by a single worker and a pool of 16 workers. The *Q*-values inferred for each colony from the single worker and the pools were similar but always lower for the latter (Table Sup IV-15), a pattern that was also observed in the Swiss samples. This is an interesting finding that deserves to be fully investigated in a larger sample size.

Discussion

The success of the numerous initiatives that are developing across Europe to protect and bring back the endangered Dark honey bee rely on molecular tools capable of accurately detecting varying levels of C-derived introgression in a time- and cost-effective manner. In many conservation programs, the breeding stock has been routinely identified through wing morphometry and, more recently, through microsatellites (Bouga *et al.* 2011). However, inferring from data on Africanized honey bees (Guzmán-Novoa *et al.*, 1994), wing morphometry is likely unable to detect low levels of C-lineage introgression into *A. m. mellifera*, a limitation that is overcome by microsatellites (Jensen *et al.*, 2005; Soland-Reckeweg *et al.*, 2008). While adoption of microsatellites represented a major step in conservation management of *A. m. mellifera* (Soland-Reckeweg *et al.*, 2008), it has been shown that a reduced number of high-graded SNPs (Muñoz *et al.*, 2015), outperform the multiallelic marker in estimating introgression (Muñoz *et al.*, 2017; Parejo *et al.*, 2018).

Here, from the 144 top-ranked SNPs, selected by their power in discriminating C from M-lineage honey bees (Muñoz *et al.*, 2015), we designed, tested and validated four assays for genotyping with the iPLEX MassARRAY system. We provide the genomic information along with the PCR and iPLEX primers for 117 high-quality SNPs multiplexed in the four assays for immediate application in genetic surveys and conservation management of *A. m. mellifera*. In addition, we provide the dataset with the genotypes for haploid and diploid individuals of *A. m. mellifera*, *A. m. carnica* and *A. m. ligustica*, which can be used by others in introgression analysis as baseline

reference populations with no need for inter-laboratory calibration (Vignal *et al.*, 2002). As opposed to microsatellites, merging of SNP databases is straightforward as there are only two alleles per locus and different platforms will provide the same allele calls. If needed, curation will only involve SNP conversion from different platforms to be on the same DNA strand, which is much simpler than trying to harmonize different microsatellite allele sizes genotyped in different laboratories.

We show that C-lineage introgression can be accurately estimated from haploid, diploid, and combined haploid and diploid datasets using either STRUCTURE (Pritchard *et al.*, 2000) or ADMIXTURE (Alexander *et al.*, 2009). These findings indicate that honey bee conservation managers can choose the software of their preference and, more importantly, can simultaneously analyse workers and drones without biasing estimates of C-lineage introgression in *A. m. mellifera* colonies.

The Assay Design software was able to combine only 127 of the 144 high-graded SNPs (Muñoz *et al.*, 2015) into four multiplexes. While the iPLEX protocol allows multiplexing up to 40 SNPs, only assay M1 (38 SNPs) approached the maximum plexing capacity. This is in part due to the relatively small size of the baseline SNP set from which the Assay Design had to work. However, the plex level of each assay can be expanded any time. By using the *Replex* option of the software, additional high-graded nuclear SNPs or even mitochondrial SNPs can be added to the customized four assays for detecting C-derived genes at both genetic compartments.

The iPLEX MassARRAY system revealed highly accurate and delivered high-quality calls for 117 of the 127 SNPs. Quality assessment was greatly facilitated by the honey bee haplodiploid system. Using the SNP calls of the drone subset, problematic SNPs were easily detected by locating genotypes erroneously typed heterozygous. Three such SNPs were consistently identified in numerous drones. While the mechanism responsible for the false allele is unclear, it is possible that gene homology is the source of miscalling at least in locus 1379-est5929. Using the 120-bp flanking region of this SNP locus, a NCBI query found a second hit with 98% similarity in the honey bee genome. The 117 SNPs were successfully genotyped in over 96% of the samples, indicating that the customized four assays and the iPLEX MassARRAY system work well in DNAs obtained from a variety of tissues with the virtually full spectrum of extraction methods routinely employed in honey bee research (Evans *et al.*, 2013).

The four combined SNP assays were able to estimate introgression with a high degree of accuracy. However, performance decreased to some extent when SNP assays were used singly and the 23-plex M4 showed the worst behaviour for most statistics. This finding is consistent with studies on other organisms which have also detected drops in accuracy when the number of SNPs is <25 (Storer *et al.*, 2012; Wilkinson *et al.*, 2012). Further assessment of the four assays (used singly or combined) at the individual level indicates that there is a greater chance of misclassifying purebred individuals as admixed than the reverse, *viz.* misclassifying admixed individuals as purebred. This result has practical implications in conservation management suggesting that it is more likely that *A. m. mellifera* genetic diversity is erroneously discarded from the breeding population than C-derived genes are maintained. At this point, simulation and empirical studies are needed to determine the best threshold criterion to separate purebreds from admixed individuals (Vähä & Primmer, 2006). While the stringent *Q*-value threshold of <0.05 arbitrarily established here for defining purebred *A. m. mellifera* may assure a more efficient purging of C-derived alleles, it may also lead to erosion of *A. m. mellifera* diversity and loss of unique gene complexes. The problem is that low diversity is particularly detrimental for honey bees because it may decrease colony resistance to brood diseases (Seeley & Tarpy, 2007) and increase genetic load at the sex locus (Page, 1980). Therefore, managers of *A. m. mellifera* conservatories need to make a trade-off between purging foreign alleles from the breeding population while minimizing the effects of reduced diversity.

Validation of the four SNP assays in an independent set of individuals, including F1 hybrids (obtained from controlled crosses purposely established for this study, as opposed to the simulated hybrids more commonly found in the literature), further confirms the resolution power of our customized SNP assays. Interestingly, the *Q*-values obtained for the F1 hybrids were in close proximity to the expected 0.50, although there was a bias towards C-derived genes as most *Q*-values were >0.50. When used singly, the SNP assays failed to correctly identify all purebred individuals and the *Q*-values were more dispersed around the expected threshold of 0.50. However, when the assays were combined, the performance increased with all purebred individuals correctly classified and the *Q*-values showing a lower dispersion around 0.50. Interestingly, despite the lower number of SNPs contained in M1+M3 (62 *vs* 117), this assay combination shows a better overall performance than the four assays together.

Sustainable conservation management requires tools capable of reliably identifying breeding colonies in a time- and cost-efficient manner. The SNP assays tested herein have a high resolution power for accurately estimating introgression, and the iPLEX MassARRAY system offers an interesting alternative for rapid and cost-effective genotyping. This system is very flexible and scalable allowing a variety of options for sample and assay throughput at a variable cost, depending on the chip format (24, 96, or 384) chosen. The 384 format, for example, allows genotyping 384 samples with a single assay at an approximate outsourced cost of 4.5€ per sample. Alternatively, this format could be used to genotype 192, 128, or 96 samples with two, three, or four assays, respectively. This option would incur in an increment of 4.5€ for any additional assay. Based on overall results, the best compromise between genotyping costs and assay accuracy is achieved when using M1+M3.

Genotyping a single microsatellite multiplex in a 96-plate format costs approximately 2.5€ per sample. Introgression proportions using microsatellites has typically been estimated from over 11 loci, which requires genotyping a minimum of two multiplexes (Bouga *et al.*, 2011; Garnery *et al.*, 1998; Jensen *et al.*, 2005; Parejo *et al.*, 2018) thereby doubling the per-sample cost. However, this charge does not include PCR and microsatellite fragment analysis. Contrary to microsatellites, outsourced SNP genotyping with the iPLEX MassArray system only requires DNA (instead of PCR products) to generate a table of genotypes ready to analyse, avoiding the hurdle of fragment analysis.

Honey bee queens mate in flight with up to 20 drones (Estoup *et al.*, 1994). This means that in areas where *A. m. mellifera* and commercial colonies are sympatric, matings may occur with drones of C-lineage ancestry originating colonies made up of subfamilies with diverse genetic backgrounds. Although population-level studies typically require genotyping a single worker per colony (Meixner *et al.* 2013), colony-level introgression estimates may require genotyping several individuals to more effectively capture the colony structure. The problem is that genotyping several workers per colony is time consuming and costly. An economical way to circumvent this issue is to genotype pools instead of individuals (Gautier *et al.*, 2013), provided that the genotyping system of choice is sensitive enough to detect low-frequency alleles.

Here, we assessed whether our customized SNPs assays and the iPLEX MassARRAY system offer a reliable alternative for pool genotyping. Both DNA and tissue pooling experiments show that the genotyping system is very sensitive as it was able to detect low frequency alleles.

Despite the small number of SNPs showing consistent amplification across experiments, introgression analysis indicates that as few as 62 SNPs (M1+M3) were able to detect highly diluted C-derived alleles. These results suggest that this system has the potential to detect C-lineage introgression in colonies with hybrid sub-families at low frequency, a scenario that might occur if drones of commercial colonies are able to accidentally enter congregation areas of conservatories.

Analysis of pool genotypes showed that miscalling was mainly due to the unequal contribution of each individual (different concentrations) and to the unbiased representation of allelic products that are present in a DNA pool, both common problems reported for DNA pools (Gautier *et al.*, 2013; Sham *et al.*, 2002). While pools constructed from equi-molar DNA concentrations would be the most correct approach to genotype a colony, pooling tissues is often the only option in conservation programs requiring screening of numerous colonies with a limited budget. Pooling tissue instead of DNA requires less time, effort and money during preparation in the laboratory and still enables detection of C-derived alleles even when most of the individuals in the pool are *A. m. mellifera*.

The introgression analysis on the samples collected throughout Europe and genotyped using the four SNP assays and the iPLEX MassARRAY system provides a rough picture of the genetic integrity of *A. m. mellifera*. This SNP survey adds to Pinto *et al.* (2014) by expanding the sampling in France, Switzerland, UK and by including *de novo* Wales and Ireland. Concordant with earlier microsatellite (Jensen *et al.*, 2005; Soland-Reckeweg *et al.*, 2008) and SNP (Parejo *et al.*, 2016; Pinto *et al.*, 2014) surveys, C-lineage introgression in *A. m. mellifera* is heterogeneous across Europe. Samples originating from conservatories were generally less introgressed than those from unprotected areas. Our previous and this SNP survey revealed that Scotland, Norway, Netherlands and now Ireland possess important pockets of pure *A. m. mellifera*. Ireland represents a particularly interesting case of *A. m. mellifera* diversity because, contrary to the other countries, the survey was performed in unprotected populations from a wide geographical area.

As this and previous studies (Jensen *et al.*, 2005; Parejo *et al.*, 2016; Pinto *et al.*, 2014; Soland-Reckeweg *et al.*, 2008) represent only partial, and in some cases biased, assessments on the status of the genetic integrity of *A. m. mellifera* across its distributional range, this novel tool now makes it possible to perform a comprehensive genetic survey in a time- and cost-efficient manner. We suggest that if the efficacy of this SNP tool is generally agreed among stakeholders the next step is for them to seek input from government agencies and/or research facilities and begin

to describe the purity of their honey bee populations on as wide a geographic area as possible in order that conservation efforts correctly and efficiently target regions of greatest concern and greatest possible reward.

Methods

Assay design

Muñoz *et al.* (2015) identified 144 highly informative SNPs for estimating C-lineage introgression in *A. m. mellifera*. The flanking regions (60 bp of either side) of these SNPs were used to design multiplexed assays with the software Assay Design 4.0 (Agena BioScience™) for genotyping using the Agena BioScience iPLEX chemistry and the MassARRAY® MALDI-TOF platform (hereafter abbreviated to iPLEX MassARRAY). The software searched for optimal areas within the 120-bp flanking regions to design forward and reverse PCR primers while constructing the different multiplexes. The maximum multiplexing capacity (40 SNPs) allowed by the iPLEX chemistry was attempted whilst preventing hairpin and dimer formation. In addition to the PCR primers, the software designed the iPLEX extension primer placed immediately adjacent to each SNP. Of the 144 SNPs, the Assay Design was able to combine 127 SNPs distributed along four multiplexed assays (see Table Sup IV-1 for sequences of the flanking regions, sequences of PCR and iPLEX reaction primers, and composition of the four multiplexes). The putative functional role of the genes marked by each SNP was identified using SNPeff 4.3 tool build (Cingolani *et al.*, 2012) and the NCBI *Apis mellifera* annotation genome version 102 (Pruitt *et al.*, 2013).

Samples and DNA extraction

A total of 464 colonies (represented by a single haploid drone, a single diploid worker, multiple workers, or pools of drones or workers; Table Sup IV-2) were sampled across Europe (Figure IV-1). The samples originated from colonies in the (i) *A. m. mellifera* (N=462) native range in Western and Northern Europe (protected and unprotected areas), (ii) *A. m. ligustica* (N=10), and *A. m. carnica* native ranges (N=10) in Southeastern Europe, (iii) introduced range of *A. m. carnica* in Switzerland (N=8), Germany (Kirchhain; N=16) and Scotland (N=3), (iv) commercial strain Buckfast from Switzerland, Scotland, and Denmark (N=11), and (v) F1 hybrid crosses performed in isolated mating stations in Denmark (N=19). Nine samples of *A. m. carnica* and seven *A. m.*

ligustica, previously genotyped using the GoldenGate® Assay in the BeadArray platform of Illumina (Pinto *et al.*, 2014), were added to the dataset to have a better representation of C-lineage.

Genomic DNA was extracted from the head, antennae, thorax (entire or half), legs, or abdomen of adults or immatures (larvae or pupae) of a single individual, multiple individuals (extracted, then pooled), or a pool of individuals (mixed tissue, then extracted) per colony in 561 samples (Table Sup IV-2). The extraction methods included phenol-chloroform, CTAB, commercial kits (Qiagen EZ1 DNA tissue kit, Omega bio-tek EZNA kit), and magnetic beads using the KingFisher™ Flex Purification System. These represent the wide array of tissues and extraction methods commonly used in honey bee research (Evans *et al.* 2013). The DNA samples were set at a concentration of 10-15 ng/µl and sent to *Instituto Gulbenkian de Ciência* (Portugal) for SNP genotyping.

SNP genotyping and quality control

A total of 573 samples (561 plus 12 DNA pools, Table Sup IV-2) were genotyped for the 127 SNP loci multiplexed in the four assays using the iPLEX chemistry and the MassARRAY® MALDI-TOF genotyping platform (Gabriel *et al.* 2009). The genotypes generated for the 573 samples (Table Sup IV-16) were subjected to quality control filters to discard SNP loci and samples with poor or inconsistent amplification. SNPs and samples with missing data >20% (Table Sup IV-1) and >30% (Table Sup IV-2), respectively, were excluded from the dataset (Table Sup IV-3).

Assessing genotyping accuracy

The genotyping accuracy was assessed on the subset of single haploid drones by (i) identifying the heterozygous SNP loci (N=171; Table Sup IV-2) and (ii) comparing the SNP calls generated for a variable number of individuals (Figure IV-2) by the iPLEX MassARRAY system with those obtained with the GoldenGate® Assay genotyped in the BeadArray platform of Illumina (N=96 individuals (Pinto *et al.*, 2014) and with the HiSeq 2500 platform of Illumina (N=32 individuals; D.H./M.A.P., unpublished data and Parejo *et al.* (2016). The SNP loci that were called heterozygous by the MassARRAY system in >10% of the drones and showed inconsistent genotypes between at least two genotyping technologies in >5% of the drones were excluded from further analysis (Tables Sup IV-1 and Sup IV-3).

Comparing approaches of introgression estimation

The clustering approach implemented by the software packages STRUCTURE (Pritchard *et al.*, 2000) and ADMIXTURE (Alexander *et al.*, 2009) has been preferred for estimating introgression (*Q*-values) in honey bees, especially of C-lineage in *A. m. mellifera* (Jensen *et al.*, 2005; Parejo *et al.*, 2016; Pinto *et al.*, 2014; Soland-Reckeweg *et al.*, 2008). While both packages handle haploid and diploid data, it is unclear whether *Q*-values are accurate for datasets combining different ploidies, a circumstance that might occur if drones and workers are required to be genotyped. To support the decision making of honey bee managers, the consistency of the *Q*-values was assessed by comparing (i) the output of ADMIXTURE and STRUCTURE generated from a subset of 112 drones, and (ii) the output of ADMIXTURE generated from workers (N=112), diploid (N=112), and combined drones and workers (N=224; Figure IV-2). Differences in *Q*-values were assessed by Mann-Whitney test.

In STRUCTURE, *Q*-values were estimated for two ancestral clusters (K=2) using the admixture ancestry and correlated allele frequency models with the unsupervised option. The initial burn-in was set to 250,000 followed by 750,000 Markov chain Monte Carlo iterations. Over 20 independent runs were performed to confirm consistency across runs. In ADMIXTURE, *Q*-values were also estimated for K=2 using 10,000 iterations in 20 independent runs. The convergence between iterations was monitored by comparing log-likelihood scores (LLS) using the default termination criterion set to stop when LLS increases by <0.0001 between iterations. CLUMPAK was used to summarize and visualize the STRUCTURE and ADMIXTURE *Q*-plots (Kopelman *et al.*, 2015).

Assessing performance of the SNP assays

The performance of the SNP assays in estimating C-lineage introgression in *A. m. mellifera* was assessed by comparing the *Q*-values inferred by them with those inferred from 2,399 million SNPs identified in WGs (D.H./M.A.P., unpublished data and Parejo *et al.* (2016). A total of 38 drones (4 *A. m. ligustica*, 7 *A. m. carnica*, 11 purebred *A. m. mellifera*, and 16 admixed *A. m. mellifera*), for which there were WG sequence data available, was used in this comparison (Figure IV-2). The performance of the four assays (individually or combined) was assessed by (i) Pearson's correlation coefficient (*r*), (ii) similarity score obtained by CLUMPAK, (iii) absolute accuracy error calculated as the absolute difference between *Q*-values inferred from the SNP assays and the 2,399 million

SNPs, (iv) mean accuracy calculated via percentage of absolute error, (v) absolute precision error calculated via standard deviation of the absolute differences, (vi) number of purebred individuals classified as admixed, and (vii) number of admixed individuals classified as purebred. Admixed individuals were defined by a threshold *Q*-value>0.05. Any individual with *Q*-value between 0 and <0.05 or >0.95 and 1 was classified as purebred *A. m. mellifera* and C-lineage (*A. m.carnica* or *A. m. ligustica*), respectively.

Validating the SNP assays

The four assays were validated and tested using an independent subset of 62 workers, including 30 *A. m. mellifera* (Endelave, Denmark), 16 *A. m. carnica* (Kirchhain, Germany), and 16 F1 hybrids obtained from crosses between *A. m. mellifera* queens, from the conservatory in Læsø, and *A. m. carnica* drones from Mandø, Denmark (Figure IV-2; Table Sup IV-10). The crosses were performed in the isolated mating station of Mandø in 2016. *Q*-values were inferred from the four assays (individually or combined) by ADMIXTURE and then compared with expected *Q*-values of >0.95 for *A. m. carnica*, <0.05 for *A. m. mellifera*, and 0.5 for the F1 hybrids.

Assessing sensitivity of the MassARRAY system in pooled DNA

Pools of tissue or DNA are a cost-efficient option for estimating introgression in organisms with a polyandrous mating system like the honey bee. However, pooling can only be adopted if the genotyping system is able to consistently detect low-frequency alleles. The sensitivity of the MassARRAY system was assessed in a dilution experiment of varying ratios of DNAs of two haploid drones: one *A. m. ligustica* and one *A. m. mellifera.* The two drones displayed the highest number of alternate alleles for the 127 highly-informative SNPs identified in a large dataset previously genotyped with the GoldenGate® Assay (Pinto *et al.*, 2014).

The experiment was performed by pooling the DNA of the two drones using volume ratios of 10:20, 5:20, 2:20, 1:20, and 0.5:20 *A. m. ligustica* to *A. m. mellifera* (Figure IV-2). The number of replicates was three for 1:20 and 0.5:20 and two for the remaining ratios, as they were nested in the higher dilution factors. The pools were genotyped for the four assays using the iPLEX MassARRAY. The genotypes generated from the pooled DNAs were compared with those expected and the number of mismatches was recorded. The expected genotypes of the pools were inferred from the SNP calls for the single drones.

The sensitivity of the genotyping system in detecting C-lineage ancestry in the pooled samples was also assessed via introgression analysis. The *Q*-values were estimated by ADMIXTURE for each DNA pool using the expected and observed genotypes for a variable number of SNPs (four assays and best assay combination, as defined by *r*).

Assessing sensitivity of the MassARRAY system in pooled tissue

The sensitivity of the MassARRAY system was further assessed in tissue pools (Figure IV-2). A total of 22 pools were constructed using varying ratios of workers (1:1, 1:2, 1:3, 1:7) of two different ancestries chosen among *A. m. mellifera*, *A. m. carnica*, Buckfast, and F1 hybrids (*A. m. mellifera* queens x *A. m. carnica* drones), as detailed in Table Sup IV-17. The DNA was extracted twice (individually and pooled) from the thorax, which had been cut in two identical portions. The DNA concentrations of individual and pooled extractions were measured using NanoDrop™ (Table Sup IV-10).

The sensitivity of the genotyping system was first assessed by comparing the SNP calls obtained for the single workers with those obtained for the pools of workers. Mismatches were counted and the error identified among the following sources: (i) pools displayed alleles uncalled in single workers and *vice versa*, (ii) SNP calls of the pools matched those of the worker with higher DNA concentration, (iii) SNP calls of the pools matched the most frequent allele, and (iv) the least frequent allele. The sensitivity of the genotyping system in detecting C-lineage ancestry in the different pools was also assessed via introgression analysis. The *Q*-values were estimated for the 22 pools from the expected and observed genotypes, for a variable number of SNPs (four assays and best assay combination), using ADMIXTURE. The expected genotypes were inferred from the calls obtained for the single workers.

Applying the SNP assays

The four assays were used to genotype in the MassARRAY platform 462 samples representing *A. m. mellifera* (N=425), *A. m. ligustica* (N=10), *A. m. carnica* (N=21), and Buckfast (N=6) from 13 European countries (Figure IV-1 and Figure IV-2). Samples of *A. m. mellifera* originated from protected (N=125) and unprotected (N=300) areas. Of the 462 samples, 415 were represented by a single individual and 47 by pooled individuals (16 pooled workers from colonies of *A. m. mellifera*, *A. m. carnica* and Buckfast; 30 pooled drones from colonies of *A. m. mellifera*; Table Sup IV-2). Additionally, a subset of four colonies (two *A. m. mellifera*, one *A. m. carnica*, and one

Buckfast) from Scotland and England was represented by both a pool of 16 workers and one individual worker. For a better C-lineage representation, nine samples of *A. m. carnica* and 7 of *A. m. ligustica* (each representing a single individual and colony), previously genotyped using the GoldenGate® Assay (Pinto *et al.*, 2014), were added to the dataset. *Q*-values were inferred from the genotypes of single and pooled samples using ADMIXTURE.

The genotype data were further examined by network analysis using the software Graphia Professional (Kajeka Ltd, Edinburgh, UK). For each sample, SNPs were scored 0 when same as reference (*A. m. carnica*), 1 for heterozygous and 2 for homozygous different to reference, i.e. representing the *A. m. mellifera* allele. Where data was missing, the SNP was scored 1.01. For ease of interpretation, the total combined score for each SNP in each sample was calculated and the SNPs reordered from the smallest score to the largest. The SNP data and associated sample metadata was loaded into Graphia and a Pearson correlation matrix was calculated comparing the profile of SNP scores for each sample. A network graph was then constructed by connecting the nodes (samples) with edges (where the correlation exceeded the threshold value r>0.27). Utilising the overlay of metadata the graph was then explored and clustered using the Markov Cluster (MCL) algorithm (Enright *et al.*, 2002) at an inflation value (which determines cluster granularity) of 1.2.

References

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*.

Bouga, M., Alaux, C., Bienkowska, M., Büchler, R., Carreck, N. L., Cauia, E., Chlebo, R., Dahle, B., Dall'Olio, R., De la Rúa, P., Gregorc, A., Ivanova, E., Kence, A., Kence, M., Kezic, N., Kiprijanovska, H., Kozmus, P., Kryger, P., Conte, Y. L., Lodesani, M., Murilhas, A. M., Siceanu, A., Soland, G., Uzunov, A., & Wilde, J. (2011). A review of methods for discrimination of honey bee populations as applied to European beekeeping. *Journal of Apicultural Research, 50*(1), 51-84. doi: 10.3896/ibra.1.50.1.06

Büchler, R., Costa, C., Hatjina, F., Andonov, S., Meixner, M. D., Conte, Y. L., Uzunov, A., Berg, S., Bienkowska, M., & Bouga, M. (2014). The influence of genetic origin and its interaction with environmental effects on the survival of *Apis mellifera* L. colonies in Europe. *Journal of Apicultural Research, 53*(2), 205-214.

Chapman, N. C., Bourgeois, A. L., Beaman, L. D., Lim, J., Harpur, B. A., Zayed, A., Allsopp, M. H., Rinderer, T. E., & Oldroyd, B. P. (2017). An abbreviated SNP panel for ancestry assignment of honeybees (*Apis mellifera*). *Apidologie, 48*(6), 776-783.

Chapman, N. C., Harpur, B. A., Lim, J., Rinderer, T. E., Allsopp, M. H., Zayed, A., & Oldroyd, B. P. (2015). A SNP test to identify Africanized honeybees via proportion of 'African' ancestry. *Molecular Ecology Resources, 15*(6), 1346-1355. doi: 10.1111/1755-0998.12411

Chen, C., Liu, Z., Pan, Q., Chen, X., Wang, H., Guo, H., Liu, S., Lu, H., Tian, S., & Li, R. (2016). Genomic analyses reveal demographic history and temperate adaptation of the newly discovered honey bee subspecies *Apis mellifera sinisxinyuan* n. ssp. *Molecular biology and evolution, 33*(5), 1337-1348.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly, 6*(2), 80-92. doi: 10.4161/fly.19695

De la Rúa, P., Jaffé, R., Muñoz, I., Serrano, J., Moritz, R. F. A., & Kraus, F. B. (2013). Conserving genetic diversity in the honeybee: Comments on Harpur et al.(2012). *Molecular Ecology, 22*(12), 3208-3210. doi: 10.1111/mec.12333

Engel, M. S. (1999). The taxonomy of recent and fossil honey bees (Hymenoptera: Apidae; Apis). *Journal of Hymenoptera Research, 8*(2).

Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research, 30*(7), 1575-1584.

Estoup, A., Solignac, M., & Cornuet, J.-M. (1994). Precise assessment of the number of patrilines and of genetic relatedness in honeybee colonies. *Proceedings of the Royal Society of London B: Biological Sciences, 258*(1351), 1-7.

Evans, J. D., Schwarz, R. S., Chen, Y. P., Budge, G., Cornman, R. S., De la Rua, P., de Miranda, J. R., Foret, S., Foster, L., & Gauthier, L. (2013). Standard methods for molecular research in *Apis mellifera*. *Journal of Apicultural Research, 52*(4), 1-54.

Gabriel, S., Ziaugra, L., & Tabbaa, D. (2009). SNP genotyping using the Sequenom MassARRAY iPLEX platform. *Current protocols in human genetics*, 2.12. 11-12.12. 16.

Garnery, L., Franck, P., Baudry, E., Vautrin, D., Cornuet, J.-M., & Solignac, M. (1998). Genetic diversity of the west European honey bee (*Apis mellifera mellifera* and *A. m. iberica*) II. Microsatellite loci. *Genetics Selection Evolution, 30*(1), S49 %@ 1297-9686.

Gautier, M., Foucaud, J., Gharbi, K., Cézard, T., Galan, M., Loiseau, A., Thomson, M., Pudlo, P., Kerdelhué, C., & Estoup, A. (2013). Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. *Molecular Ecology, 22*(14), 3766-3779.

Guzmán-Novoa, E., Page Jr, R. E., & Fondrk, M. K. (1994). Morphometric techniques do not detect intermediate and low levels of Africanization in honey bee (Hymenoptera: Apidae) colonies. *Annals of the Entomological Society of America, 87*(5), 507-515.

Harpur, B. A., Kent, C. F., Molodtsova, D., Lebon, J. M., Alqarni, A. S., Owayss, A. A., & Zayed, A. (2014). Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. *Proceedings of the National Academy of Sciences of the United States of America, 111*(7), 2614-2619. doi: 10.1073/pnas.1315506111

Jensen, A. B., Palmer, K. A., Boomsma, J. J., & Pedersen, B. V. (2005). Varying degrees of *Apis mellifera ligustica* introgression in protected populations of the black honeybee, *Apis mellifera mellifera*, in northwest Europe. *Molecular Ecology, 14*(1), 93-106.

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., & Mayrose, I. (2015). Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Mol Ecol Resour, 15*(5), 1179-1191. doi: 10.1111/1755-0998.12387

Meixner, M. D., Francis, R. M., Gajda, A., Kryger, P., Andonov, S., Uzunov, A., Topolska, G., Costa, C., Amiri, E., & Berg, S. (2014). Occurrence of parasites and pathogens in honey bee colonies used in a European genotype-environment interactions experiment. *Journal of Apicultural Research, 53*(2), 215-229.

Meixner, M. D., Kryger, P., & Costa, C. (2015). Effects of genotype, environment, and their interactions on honey bee health in Europe. *Current Opinion in Insect Science, 10*, 177-184. doi: 10.1016/j.cois.2015.05.010

Meixner, M. D., Leta, M. A., Koeniger, N., & Fuchs, S. (2011). The honey bees of Ethiopia represent a new subspecies of *Apis mellifera- Apis mellifera simensis* n. ssp. . *Apidologie, 42*, 425-437. doi: 10.1007/s13592-011-0007-y

Meixner, M. D., Pinto, M. A., Bouga, M., Kryger, P., Ivanova, E., & Fuchs, S. (2013). Standard methods for characterising subspecies and ecotypes of *Apis mellifera*. *Journal of Apicultural Research, 52*(4), 1-28.

Muñoz, I., Henriques, D., Jara, L., Johnston, J. S., Chávez-Galarza, J., De La Rúa, P., & Pinto, M. A. (2017). SNPs selected by information content outperform randomly selected microsatellite loci for delineating genetic identification and introgression in the endangered Dark European honeybee (*Apis mellifera mellifera*). *Molecular Ecology Resources, 17*(4), 783-795.

Muñoz, I., Henriques, D., Johnston, J. S., Chávez-Galarza, J., Kryger, P., & Pinto, M. A. (2015). Reduced SNP panels for genetic identification and introgression analysis in the Dark honey bee (*Apis mellifera mellifera*). *PLoS One, 10*(4), e0124365. doi: 10.1371/journal.pone.0124365

Page, R. E. (1980). The evolution of multiple mating behavior by honey bee queens (*Apis mellifera L.*). *Genetics, 96*(1), 263-273.

Parejo, M., Henriques, D., Pinto, M. A., , S.-R., G, & Neuditschkoa, M. (2018). Empirical comparison of microsatellite and SNP markers to estimate introgression in *Apis mellifera mellifera*. *Journal of Apicultural Research, submitted*.

Parejo, M., Wragg, D., Gauthier, L., Vignal, A., Neumann, P., & Neuditschko, M. (2016). Using Whole-Genome Sequence information to foster conservation efforts for the european Dark honey bee, *Apis mellifera mellifera*. *Frontiers in Ecology and Evolution, 4*(140). doi: 10.3389/fevo.2016.00140

Pinto, M. A., Henriques, D., Chávez-Galarza, J., Kryger, P., Garnery, L., van der Zee, R., Dahle, B., Soland-Reckeweg, G., De la Rúa, P., Dall' Olio, R., Carreck, N., & Johnston, J. S. (2014). Genetic integrity of the Dark European honey bee (*Apis mellifera mellifera*) from protected populations: a genome-wide assessment using SNPs and mtDNA sequence data. *Journal of Apicultural Research, 53*(2), 269-278. doi: 10.3896/ibra.1.53.2.08

Potts, S. G., Biesmeijer, J. C., Kremen, C., Neumann, P., Schweiger, O., & Kunin, W. E. (2010). Global pollinator declines: trends, impacts and drivers. *Trends in Ecology & Evolution, 25*(6), 345-353. doi: 10.1016/j.tree.2010.01.007

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics, 155*(2), 945-959.

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., & McGarvey, K. M. (2013). RefSeq: an update on mammalian reference sequences. *Nucleic Acids Research, 42*(D1), D756-D763. doi: 10.1093/nar/gkt1114

Ruttner, F. (1988). *Biogeography and Taxonomy of Honey Bees.*

Seeley, T. D., & Tarpy, D. R. (2007). Queen promiscuity lowers disease within honeybee colonies. *Proceedings of the Royal Society of London B: Biological Sciences, 274*(1606), 67-72.

Sham, P., Bader, J. S., Craig, I., O'Donovan, M., & Owen, M. (2002). DNA Pooling: a tool for large-scale association studies. *Nature Reviews Genetics, 3*(11), 862-871.

Sheppard, W. S., & Meixner, M. D. (2003). *Apis mellifera pomonella*, a new honey bee subspecies from Central Asia. *Apidologie, 34*, 367–375. doi: 10.1051/apido:2003037

Soland-Reckeweg, G., Heckel, G., Neumann, P., Fluri, P., & Excoffier, L. (2008). Gene flow in admixed populations and implications for the conservation of the Western honeybee, *Apis mellifera*. *Journal of Insect Conservation, 13*(3), 317. doi: 10.1007/s10841-008-9175-0

Storer, C. G., Pascal, C. E., Roberts, S. B., Templin, W. D., Seeb, L. W., & Seeb, J. E. (2012). Rank and order: evaluating the performance of SNPs for individual assignment in a non-model organism. *PLoS One, 7*(11), e49018.

Vähä, J.-P., & Primmer, C. R. (2006). Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. *Molecular Ecology, 15*(1), 63-72.

vanEngelsdorp, D., & Meixner, M. D. (2010). A historical review of managed honey bee populations in Europe and the United States and the factors that may affect them. *Journal of Invertebrate Pathology, 103, Supplement*, S80-S95. doi: 10.1016/j.jip.2009.06.011

Vignal, A., Milan, D., SanCristobal, M., & Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution, 34*(3), 275.

Weinstock, G. M., Robinson, G. E., Gibbs, R. A., Worley, K. C., Evans, J. D., Maleszka, R., Robertson, H. M., Weaver, D. B., Beye, M., Bork, P., Elsik, C. G., Hartfelder, K., Hunt, G. J., Zdobnov, E. M., Amdam, G. V., Bitondi, M. M. G., Collins, A. M., Cristino, A. S., Lattorff, H. M. G., Lobo, C. H., Moritz, R. F. A., Nunes, F. M. F., Page Jr, R. E., Simões, Z. L. P., Wheeler, D., Carninci, P., Fukuda, S., Hayashizaki, Y., Kai, C., Kawai, J., Sakazume, N., Sasaki, D., Tagami, M., Albert, S., Baggerman, G., Beggs, K. T., Bloch, G., Cazzamali, G., Cohen, M., Drapeau, M. D., Eisenhardt, D., Emore, C., Ewing, M. A., Fahrbach, S. E., Forêt, S., Grimmelikhuijzen, C. J. P., Hauser, F., Hummon, A. B., Huybrechts, J., Jones, A. K., Kadowaki, T., Kaplan, N., Kucharski, R., Leboulle, G., Linial, M., Littleton, J. T., Mercer, A. R., Richmond, T. A., Rodriguez-Zas, S. L., Rubin, E. B., Sattelle, D. B., Schlipalius, D., Schoofs, L., Shemesh, Y., Sweedler, J. V., Velarde, R., Verleyen, P., Vierstraete, E., Williamson, M. R., Ament, S. A., Brown, S. J., Corona, M., Dearden, P. K., Dunn, W. A., Elekonich, M. M., Fujiyuki, T., Gattermeier, I., Gempe, T., Hasselmann, M., Kadowaki, T., Kage, E., Kamikouchi, A., Kubo, T., Kucharski, R., Kunieda, T., Lorenzen, M., Milshina, N. V., Morioka, M., Ohashi, K., Overbeek, R., Ross, C. A., Schioett, M., Shippy, T., Takeuchi, H., Toth, A. L., Willis, J. H., Wilson, M. J., Gordon, K. H. J., Letunic, I., Hackett, K., Peterson, J., Felsenfeld, A., Guyer, M., Solignac, M., Agarwala, R., Cornuet, J. M., Monnerot, M., Mougel, F., Reese, J. T., Schlipalius, D., Vautrin, D., Gillespie, J. J., Cannone, J. J., Gutell, R. R., Johnston, J. S., Eisen, M. B., Iyer, V. N., Iyer, V., Kosarev, P., Mackey, A. J., Solovyev, V., Souvorov, A., Aronstein, K. A., Bilikova, K., Chen, Y. P., Clark, A. G., Decanini, L. I., Gelbart, W. M., Hetru, C., Hultmark, D., Imler, J. L., Jiang, H., Kanost, M., Kimura, K., Lazzaro, B. P., Lopez, D. L., Simuth, J., Thompson, G. J., Zou, Z., De Jong, P., Sodergren, E., Csurös, M., Milosavljevic, A., Osoegawa, K., Richards, S., Shu, C. L., Duret, L., Elhaik, E., Graur, D., Anzola, J. M., Campbell, K. S., Childs, K. L., Collinge, D., Crosby, M. A., Dickens, C. M., Grametes, L. S., Grozinger, C. M., Jones, P. L., Jorda, M., Ling, X., Matthews, B. B., Miller, J., Mizzen, C., Peinado, M. A., Reid, J. G., Russo, S. M., Schroeder, A. J., St. Pierre, S. E., Wang, Y., Zhou, P., Jiang, H., Kitts, P., Ruef, B., Venkatraman, A., Zhang, L., Aquino-Perez, G., Whitfield, C. W., Behura, S. K., Berlocher, S. H., Sheppard, W. S., Smith, D. R., Suarez, A. V.,

Tsutsui, N. D., Wei, X., Wheeler, D., Havlak, P., Li, B., Liu, Y., Jovilet, A., Lee, S., Nazareth, L. V., Pu, L. L., Thorn, R., Stolc, V., Newman, T., Samanta, M., Tongprasit, W. A., Claudianos, C., Berenbaum, M. R., Biswas, S., De Graaf, D. C., Feyereisen, R., Johnson, R. M., Oakeshott, J. G., Ranson, H., Schuler, M. A., Muzny, D., Chacko, J., Davis, C., Dinh, H., Gill, R., Hernandez, J., Hines, S., Hume, J., Jackson, L., Kovar, C., Lewis, L., Miner, G., Morgan, M., Nguyen, N., Okwuonu, G., Paul, H., Santibanez, J., Savery, G., Svatek, A., Villasana, D., & Wright, R. (2006). Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature, 443*(7114), 931-949. doi: 10.1038/nature05260

Wilkinson, S., Archibald, A. L., Haley, C. S., Megens, H.-J., Crooijmans, R. P., Groenen, M. A., Wiener, P., & Ogden, R. (2012). Development of a genetic tool for product regulation in the diverse British pig breed market. *BMC Genomics, 13*(1), 580.

Acknowledgements

We are deeply indebted to João Costa (Instituto Gulbenkian Ciência, Oeiras, Portugal), for designing the multiplexes and SNP genotyping. José Rufino provided computational resources at the Polytechnic Institute of Bragança, Portugal. DH was supported by a PhD scholarship (SFRH/BD/84195/2012) from the Portuguese Science Foundation (FCT). KAB receives a PhD fellowship from the Irish Research Council. MP was supported by the Swiss Federal Office for Agriculture FOAG and the Fondation Sur-la-Croix, Basel. IM was supported by Saavedra Fajardo fellowship from the Fundación Séneca (20036/SF/16). MAP is a member of and receives support from the COST Action FA1307 (SUPER-B). Funding for genotyping of Irish honey bees was gratefully received from the Eva Crane Trust, the Native Irish Honey bee Society and the Department of Agriculture, Food and the Marine (16/GR/09). MB and TCF are funded by an Institute Strategic Grant from the Biotechnology and Biological Sciences Research Council (BBSRC) (BB/JO1446X/1). Financial support for this research was provided to MAP and LG by 2013-2014 BiodivERsA/FACCE-JPI joint call for research proposals, with the national funders FCT (Portugal), "Agence Nationale de la Recherche" (France), and "Ministério de Economia y Competividade" (Spain).

Data accessibility

A. m. carnica and A. m. mellifera whole-genome sequence data is deposited at the ENA (www.ebi.ac.uk/ena) under study accession number PRJEB16533.

Chapter V.

Local adaptation in the Iberian honey bee, *Apis mellifera iberiensis*: insights from whole genomes.

Dora Henriques, Andreas Wallberg, Julio Chávez-Galarza, J. Spencer Johnston, Matthew T. Webster, M. Alice Pinto

Abstract

Whole-genome scans provide insights into the molecular basis of adaptation and these will help predictions on how organisms will respond to increasing selection pressures related with climate change. Here, we resequenced 87 whole genomes of Iberian honey bees and used conceptually different selection methods (Samβada, LFMM, PCAdapt, iHs) together with *in silico* protein modelling to search for signatures of selection along climatic gradients in Iberia. We found 830 outlier SNPs most of which associated with precipitation, cloud cover, latitude and longitude. Over 90.2% of outlier SNPs lay outside exons and those located in the immediate vicinity of exons and in UTRs were found enriched. Enrichment was also detected in exonic SNPs and *in silico* protein modelling suggests that several non-synonymous SNPs are likely direct targets of selection as they lead to amino acid replacements in functionally important sites of proteins. We identified genomic signatures of local adaptation in 181 genes, many of which putatively implicated in fitness-related functions such as reproduction, immunity, olfaction, lipid biosynthesis and circadian clock; Yet, selection signatures in genes involved in transcription regulation and enrichment of SNPs laying outside of exons suggest that regulatory mechanisms are a major driver of adaptive change in the Iberian honey bee.

Keywords: Local adaptation, Iberian Honey Bee, Environmental Data.

Introduction

In the current context of a global human-mediated environmental crisis, the long-standing goal of uncovering the genetic basis of adaptation has never been so important because it will enable predictions on how organisms will respond to a rapidly changing world, which, in turn, will help design mitigating strategies. Recent technological advances allow for major steps towards that goal. Increasingly powerful high-throughput sequencing and computational technologies, coupled with increasingly sophisticated analytical tools, have changed the scale of analysis from limited genomic regions and few loci to whole genomes, allowing thereby detection of signatures of selection at an unprecedented resolution and depth.

Most genome-wide analytical tools detect selection by searching for unusual patterns of genetic variation positing that population demographic history affects variation across all loci while natural selection operates at specific loci (Biswas & Akey, 2006; Guillot *et al.*, 2014; Luikart *et al.*, 2003; Nielsen *et al.*, 2005). Known as outlier tests, selection footprints are sought by scanning genomes using a population-based differentiation measure such as F_{ST} (Excoffier *et al.*, 2009; Foll & Gaggiotti, 2008) or by an individual-based approach centred on Bayesian factor models (Duforet-Frebourg *et al.*, 2014). A peculiarity of outlier tests is that they may uncover loci subject to any type of selective pressure. This contrasts with another class of increasingly popular analytical tools, known as genetic-environment association (GEA) methods, which identify selection by finding strong associations between genetic and environmental data (Coop *et al.*, 2010; Frichot *et al.*, 2015; Hoban *et al.*, 2016; Prunier *et al.*, 2011; Stucki *et al.*, 2014). By uncovering loci that are directly or indirectly correlated with the environmental factors, GEA methods can identify the pressure driving local adaptation (Joost *et al.*, 2007; Lv *et al.*, 2014; MacCallum & Hill, 2006). A major drawback of both classes of tools is that demographic processes and complex spatial structuring may create patterns resembling selection, leading to false positives (Forester *et al.*, 2016; Jensen *et al.*, 2005; Manel *et al.*, 2016). An emerging approach controls for population structure using latent factors estimated considering the statistical model and the data simultaneously. This approach has been incorporated into some outlier tests (e. g. the Bayesian factor model of PCAdapt; (Duforet-Frebourg *et al.*, 2014) and GEA methods (e. g. latent factor mixed model, LFMM (Frichot *et al.*, 2015).

Studies using whole-genome scans have employed these analytical tools to identify hundreds of regions under selection in many model and non-model organisms (Benjelloun *et al.*,

2015; Božičević *et al.*, 2016; Garud *et al.*, 2015; Kang *et al.*, 2016; Ko *et al.*, 2014; Lai *et al.*, 2016; Makinen *et al.*, 2015; Sun *et al.*, 2014; Triska *et al.*, 2015; Wang *et al.*, 2016; Xia *et al.*, 2015). This study further contributes to the rapidly growing list of organisms by helping uncover genetic pathways underlying local adaptation of one of the most diverse and evolutionarily complex honey bee subspecies, the Iberian honey bee (hereafter IHB), *Apis mellifera iberiensis*.

The honey bee (*Apis mellifera* L.) evolved into 31 currently recognized subspecies (Chen *et al.*, 2016; Engel, 1998; Meixner *et al.*, 2011; Ruttner, 1988; Sheppard & Meixner, 2003), which have been grouped into four main evolutionary lineages: Northern and Western European, M; Southeastern European, C; African, A; and Middle Eastern, O (Ruttner, 1988). In this wide range of diversity, the M-lineage IHB is one of the most intriguing subspecies, exhibiting complex patterns of clinal variation as have many other organisms that evolved in the Iberian glacial refuge (reviewed by Weiss and Ferrand (2007). Genetic surveys of the IHB have suggested that while evolutionarily neutral processes have played an important role in shaping the sharp northeastern-southwestern Iberian cline (Cánovas *et al.*, 2011; Chávez-Galarza *et al.*, 2015; Franck *et al.*, 1998; Miguel *et al.*, 2011; Miguel *et al.*, 2007), selection is a force that cannot be ignored (Chávez-Galarza *et al.*, 2013). Iberia possesses high physiographic complexity, with several large mountain ranges, and due to its geographical position is under the influence of both the North Atlantic and the Mediterranean Sea. These features have shaped a diverse array of climates (including desert, Mediterranean, Alpine, and Atlantic) and plant communities with variable flowering peaks to which the IHB had to adapt.

A previous selection scan of the IHB using an array of 383 SNPs identified 34 putatively adaptive SNPs located in genes involved in vision, xenobiotic detoxification, and innate immune response (Chávez-Galarza *et al.*, 2013). However, the 383 SNPs were widely spaced, and given the unusually high recombination rate in honey bees (Beye *et al.*, 2006; Wang *et al.*, 2016) genomic regions important in local adaptation have certainly been missed, as suggested by whole-genome studies of other subspecies (Chen *et al.*, 2016; Fuller *et al.*, 2015; Harpur *et al.*, 2014; Mikheyev *et al.*, 2015; Wallberg *et al.*, 2014; Wallberg *et al.*, 2016). In the present study, we employed a combination of outlier and GEA methods to identify genome-wide signatures of selection from 87 whole-genome sequences, thereby expanding the SNP-array scan of Chávez-Galarza *et al.* (2013) by over 3 orders of magnitude (3367 fold). We approached local adaption in the IHB by addressing the following questions: Does adaptation arise from mutations that change amino acids or regulate

gene expression? Which genes are responsible for adaptation to different environments? Which environmental factors might act as selective pressures in IHBs? In answering these questions, major insights will be gained toward understanding the genetic pathways used by the IHB to adapt to the broad range of Iberian environments.

Methods

Sampling

A total of 87 haploid *A. m. iberiensis* males were collected in 2010 from 16 sampling sites distributed across three north-south transects: one along the Atlantic coast (AT: N=31), one along the centre (CT: N=33), and another along the Mediterranean coast (MT: N=23; see Chávez-Galarza *et al.*, 2013 for further sampling details). The sites cover a wide variety of climates ranging from the semi-arid in the southeastern part of Iberia to oceanic in the northwestern part (Figure V-1). Each of the 87 individuals represents a single colony and apiary.

Figure V-1 - Location of sampling sites distributed across the three transects in the Iberian Peninsula: Atlantic (AT; N=31), Central (CT; N=33), and Mediterranean (MT, N=23). Each dot represents a single colony and apiary. Sampling site codes (AT1 to MT6) correspond to those reported by Chávez-Galarza et al. (2013).

Environmental Variables

Geographical coordinates, recorded for each apiary using a global positioning system (GPS), were used to obtain seven environmental variables from publicly available databases (WorldClim, Climatic Research Unit, OPENEI): precipitation (prec), minimum temperature (tmin), mean temperature (tmean), maximum temperature (tmax), cloud cover (cld), relative humidity (rh), and insolation (ins). These variables were integrated into a geographic information system (ArcGIS 9.3 from ESRI) to extract yearly, seasonal and monthly data. Arabic numerals appended to each environmental variable designate the month for which the variable was obtained; for example, prec5 refers to precipitation in May. In addition to climate, land cover was described for each apiary by calculating the percentage of level 3 land cover classes (Heymann *et al.*, 1994) within a 3 km radius circular area (for further details, see Chávez-Galarza *et al.*, 2013).

To prevent potential problems caused by non-independency, environmental variables were first organized into orthogonal vectors by performing a principal component analysis (PCA) using the *ade4* package (Thioulouse *et al.*, 1997). The strong correlation between many of the environmental variables in each vector means that they share a substantial amount of information and the relative importance of each variable is difficult to assess. Accordingly, variables that were correlated at $|r| > 0.8$ (Manel *et al.*, 2010) were removed from the dataset. From an initial set of 123 environmental variables, 13 uncorrelated variables together with longitude (long) and latitude (lat), which are proxies for climatic diversity, were retained for further analysis (Tables Sup V-1, Sup V-2 and Figure Sup V-1). Each retained variable is representative of a group of highly correlated variables, as listed in the Table Sup V-2. The two largest groups comprise 33 variables; one of them is prec5, which represents a wide array of variables, including precipitation, temperature, cloud and insolation; the other group is tmin6, which only represents temperature. Latitude is correlated with 27 variables, most of which represent insolation (ins), but also spring and summer precipitation (Table Sup V-2).

Whole-Genome Sequencing and Filtering

Whole genome sequencing (WGS) was performed using the Illumina HiSeq 2500 platform, which produced a mean coverage of 11X, ranging from 3X to 23X (Table Sup V-3). Sequencing libraries were generated using Illumina TruSeq™ Sample Preparation kits. The 2X150 paired end sequence

reads were mapped against the reference honey bee genome Amel_4.5 using the Burrows-Wheeler Aligner (BWA; Li & Durbin, 2010).

To improve the read mapping quality, PCR duplicates were identified and marked using Picard (http://broadinstitute.github.io/picard/) and realignment around indels was performed to correct inconsistently mapped reads using the Genome Analysis Toolkit (GATK; DePristo *et al.*, 2011). To facilitate parallelization, the reads were split per chromosome using SAMtools (http://samtools.sourceforge.net/) and the readgroups information was modified with Picard. Bayesian population-based SNP calling was implemented using FreeBayes (Garrison & Marth, 2012) across the 87 samples. To reduce poor mapping and spurious heterozygous positions, SNPs were removed that (1) had more than two alleles, (2) showed a quality score <50, (3) were present in less than 61 samples (70%), and (4) exhibited very high (>3000) or very low (<87) read depth (Table Sup V-4). Haploid male data were intentionally misspecified to be diploid in the FreeBayes SNP calling process. Positions that showed more than 10 individuals as heterozygous were discarded, as they were unlikely to represent true SNPs. Missing genotypes were imputed by IMPUTE2 (Howie *et al.*, 2009; Purcell *et al.*, 2007). SNPs showing a minor allele frequency (MAF) <0.05 were removed from the dataset using PLINK (Purcell *et al.*, 2007).

Genomic Information

Annotation information was obtained for all SNPs, including physical position, strand orientation and SNP functional state (non-synonymous, synonymous, intron or exon UTR, or intergenic regions), using the reference genome Amel_4.5, the Official Gene Set 3.2 (BEEBASE), and the Entrez Gene of NCBI. To have a complete functional annotation of each candidate gene, putative Gene Ontology classifications were obtained based on homology to *Drosophila melanogaster* using FLYBASE. The sequence alignments spanned at least 50 peptides with an e-score of 0.5 to assign orthologs. Approximately 7,103 *D. melanogaster* genes were linked to honey bee orthologs using these criteria. DAVID v.8.0 (the Database for Annotation, Visualization and Integrated Discovery) was accessed to determine if candidate genes were enriched for a specific functional annotation (Huang *et al.*, 2009). Genes were considered as candidates for selection if they were tagged by one or more SNP outliers laying in exons, introns, or UTRs.

Population Structure

Population structure was inferred from two different approaches: PCAdapt fast (Duforet-Frebourg *et al.*, 2014; Duforet-Frebourg *et al.*, 2015) and sNMF (Frichot *et al.*, 2014). PCAdapt fast infers population structure using latent factors or scores. The approach sNMF is based on sparse non-negative matrix factorization to estimate the genetic ancestry components for each individual (Frichot *et al.*, 2014). Ten runs were performed in sNMF with *alpha*=8 for each K value (1 to 10). Cross-entropy was used to guide the choice of the number of ancestral populations. To summarize and visualize the sNMF outputs, Q-plots were post-processed online with CLUMPAK (Kopelman *et al.*, 2015). The results from the PCAdapt fast and sNMF were used to create latent factors in models (see the section below for further details).

Searches for Signatures of Local Adaptation

The whole genomes of the 87 IHBs were scanned for selection signals using three conceptually different methods (Samβada, LFMM and PCAdapt) and two datasets (a genomic dataset and a combined genomic and environmental dataset). Outlier SNPs detected by at least two methods were further examined using the haplotype-based method iHS and protein modelling. Implementation of conceptually diverse approaches allows identification of potential false positives; by cross-validating outlier SNPs there is stronger evidence for selection (de Villemereuil *et al.*, 2014; Vasemagi & Primmer, 2005).

Genetic-Environment Association Methods

Two GEA methods were employed to search for signatures of local adaptation. One implements mixed models (LFMM) and the other a logistic regression model (Sam*β*ada). LFMM uses an MCMC algorithm for regression analysis that models random effects, such as population history and isolation-by-distance, as unobserved (latent) factors (Frichot *et al.*, 2013). This approach has proven to be efficient in screening genomes for signatures of local adaptation, performing well in cases of weak selection, complex hierarchical structure and polygenic selection (de Villemereuil *et al.*, 2014; Lotterhos & Whitlock, 2015; Lv *et al.*, 2014; Rellstab *et al.*, 2015; Zueva *et al.*, 2014). The program was run using 50,000 iterations and a burn-in of 25,000. Based on the ancestry estimates previously obtained with sNMF (Frichot *et al.*, 2014) and PCAdapt (Duforet-Frebourg *et al.*, 2014), two latent factors were assumed. Since LFMM uses a stochastic algorithm, five runs with different seeds were performed. To increase the power of the LFMM test, the median z-score

and adjustment of P-value were calculated. Significance was assessed using the false discovery rate (FDR) procedure (Benjamini & Hochberg, 1995; Storey & Tibshirani, 2003).

The other GEA method Sam*β*ada is a spatial approach that uses univariate logistic regression models to identify locus-environment associations and at the same time measure spatial autocorrelation (Joost *et al.*, 2007; Stucki *et al.*, 2014; Wiegand *et al.*, 2015). Sam*β*ada was run for each of the 15 environmental variables. The analysis included global and local autocorrelation using a weighting factor based on the 25 nearest neighbours. The univariate models were ranked according to the Wald statistic. From the Wald score, the P-values for each of the 15 environmental variables were calculated and only the SNPs with a P-value <0.0001 were considered as outliers.

Frequency-Based Method – PCAdapt fast

The frequency-based PCAdapt fast approach (Duforet-Frebourg *et al.*, 2014; Duforet-Frebourg *et al.*, 2015) implements a genome scan to detect genes involved in local adaptation by taking into consideration population structure. PCAdapt fast infers population structure using latent factors or scores, and searches for loci that are atypically related to population structure measured by factor analysis (h). To calculate the best K, PCAdapt fast was run with K=10. Given that the best K was 2, as determined by eigenvalues, the software was run for the second time to infer the loci under selection for K=2. The latent factors, which describe population structure, were plotted in the first two PCA components (PC1 and PC2).

Haplotype-Based Method – iHS

The integrated haplotype score method, iHS, measures the strength of evidence for selection acting at or near a given SNP, tracking the decay of haplotype homozygosity for ancestral and derived haplotypes extending from a tested core (Szpiech & Hernandez, 2014; Voight *et al.*, 2006). The |iHS| values were estimated for candidate SNPs detected by at least two of the three previous methods using the Selscan package (Szpiech & Hernandez, 2014) with default parameters: –max-extend 1,000,000 (maximum EHH extension in bp), –max-gap 200,000 (maximum gap allowed between two SNPs in bp), –cutoff 0.05 (EHH decay cutoff). The script NORM, provided by Selscan, was implemented to frequency-normalize the output using the default parameter –bins 100 (number of frequency bins) over all chromosomes (Schlamp *et al.*, 2016; Triska *et al.*, 2015). Values of |iHS|>2 are indicative of strong signals of recent positive selection (Voight *et al.*, 2006).

In Silico Analysis of 3D Protein Structure

To understand how SNPs causing amino acid changes could interfere with protein function, the 3D structure and stability were predicted for the different variants. Structures of related proteins were searched for on Phyre2 (Kelley *et al.*, 2015) and the SWISS-MODEL servers (Biasini *et al.*, 2014). The five best matches were aligned and compared with a reference protein using MEGA7 (Kumar *et al.*, 2016); the structure with the best similarity and coverage was downloaded from the RCSB Protein Data Bank. The 3D structures of reference proteins and variants were modelled using SWISS-MODEL. FoldX (Schymkowitz *et al.*, 2005) and the 3Drefine (Bhattacharya *et al.*, 2016) servers were used to refine the 3D structures. Protein stability of each variant was predicted using the Gibbs-free energy ($\Delta\Delta G$) calculated with the FoldX software. The minimum energy required for stable structure was estimated using GROMOS96 implemented in Swiss Pdb-viewer software (Guex & Peitsch, 1997). Root-Median-Square Deviations (RMSD) between the reference protein and each variant were estimated using TM-score (Zhang & Skolnick, 2005). The 3D predicted protein structures were visualized in Pymol 0.99 (PyMOL Molecular Graphics System).

Results

A total of 1,289,449 SNPs were retained after the filtering process and using a MAF>0.05. Of these, 670,738 SNPs were located in intergenic regions (120,301 in intergenic regions <2 Kb of exons, 37,058 in intergenic regions <1 Kb upstream of exons), 557,334 in introns (23,092 <50 bp of exons), 18,841 in UTRs, and 42,536 in exons (Table Sup V-5). The average physical distance between SNPs was 170.262 bp varying between 1 bp and 136,266 bp (Figure Sup V-2).

Population Structure

Population structure and demographic history can create genomic patterns that mimic selection. Accordingly, population structure was analysed to prevent discovery of false positives (Forester *et al.*, 2016; Jensen *et al.*, 2005; Manel *et al.*, 2016; Meirmans, 2012). The genetic structure was inferred from the 1,289,449 SNPs with sNMF and PCAdapt, which identified one and two optimal number of clusters (K), respectively (Figure Sup V-3). Incongruent optimal K values can be obtained by different methods (Campana *et al.*, 2011), especially in the presence of low levels of population differentiation (Waples & Gaggiotti, 2006), which is the case of the IHB with a global F_{ST}=0.021. Despite the optimal K=1 obtained by sNMF, further partitioning of the genome revealed a clinal pattern of variation, with the northern populations of the central and Mediterranean

transects carrying an important genomic component assigned to the orange cluster (0.65 for K=2, Figure V-2a). This component decreased gradually towards the south and is absent in most Atlantic populations. Greater K values (K≥3) highlight the distinctiveness of the Atlantic populations. The clinal pattern of variation in the Mediterranean (MT) and central populations (CT) is captured by PC2, with the distinct Atlantic populations (AT) captured by the PC1 generated by PCAdapt fast (Figure V-2b). These genome-wide results confirm the Iberian cline captured by the 383 SNPs, and the claim that modern beekeeping has not disrupted the natural pattern of variation in IHBs (Chávez-Galarza *et al.*, 2015).

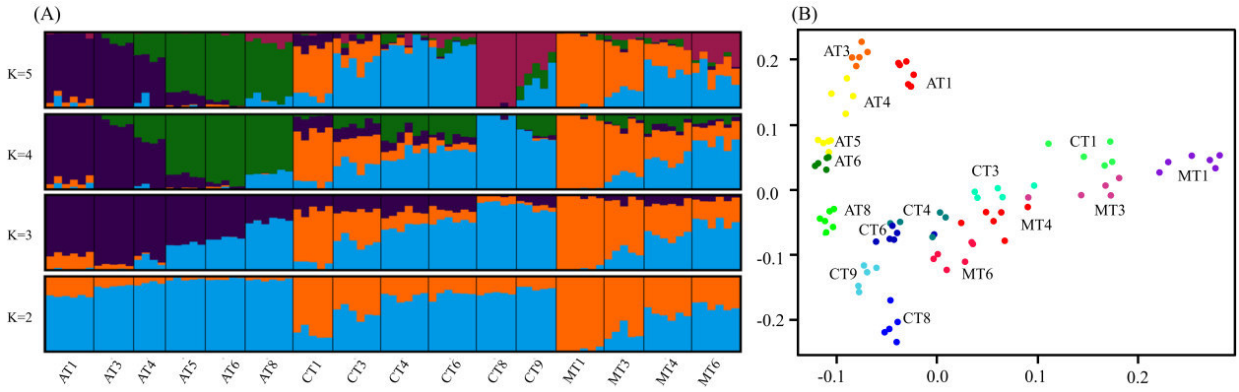


Figure V-2 - Population structure of *A. m. iberiensis* (A) estimated by sNMF from K=2 to K=5. The 16 sampling sites are arranged from north (AT1, CT1, MT1) to south (AT8, CT9, MT6) in each of the three transects. Plots represent each of the 87 individuals by a vertical bar partitioned into coloured segments (clusters) corresponding to membership proportions (Y-axis: 0-1) in each cluster. Vertical black lines separate the 16 sampling sites. (B) Score plot displaying the latent factors of each individual honey bee in PC1 and PC2 for K=2. Each colour represents a different population.

Signatures of Local Adaptation

Genetic-Environment Associations (GEA)

A total of 38,683,470 univariate models (1,289,449 SNPs x 2 alleles x 15 environmental variables) were processed by Samβada with 1,574 SNPs retained in the top-ranked 4,290 models (P-value<0.0001; Table Sup V-6). Waldscore statistics associated with each model varied between 26.1 (P-value=3.16x10-7) and 15.1 (P-value=9.98×10^{-5}), and the most frequently associated environmental variables were long (1,374, 32%), lat (828, 19%), prec5 (636, 15%), and prec1 (496, 12%). The 10 top-ranked models (Waldscore>25) identified 5 SNPs, which were located in genes GB40077 (1 SNP), GB54460 (1 SNP), GB46620 (2 SNPs) and GB48105 (1 SNP). The strongest SNP (Waldscore=26.1) was non-synonymous tagging GB40077, whereas the other four

were located in introns in the immediate vicinity of exons (between 8 and 225 bp; Table Sup V-6). Three of the 5 SNPs located in genes GB40077, GB54460, GB48105 exhibited strong associations with long and two (both located in gene GB46620) with prec5.

A total of 1,416 (FDR<0.05), 360 (FDR<0.02), and 220 SNPs (FDR<0.01) were identified by the LFMM method (Table Sup V-7 and Figure Sup V-4). The strongest 21 SNPs (defined by a cut-off level of –log10(q-value) > 4) were located in introns (11 in GB46620, 3 in GB43005, 1 in GB4810, 1 in GB54460), 1 in UTRs (GB48105), and 3 in exons (1 non-synonymous in GB46620, 1 non-synonymous in GB40077, 1 synonymous in GB48105). A single SNP mapped to an intergenic region, although close to a gene (207 bp upstream of GB46621). Most SNPs were associated with lat (11 in GB46620, 3 in GB43005, 1 in GB46621) and/or prec5 (12 in GB46620, 1 in GB46621). The variables prec1 and long were associated with only 3 (GB48105) and 2 (1 in GB40077, 1 in GB54460) SNPs, respectively.

A total of 814 SNPs overlapped between Samβada and LFMM (Tables Sup V-8 and Sup V-9). These SNPs mapped to 179 genes and 144 intergenic regions. The variables lat and prec5 showed the greatest number of associated SNPs (350 and 308, respectively; Table V-1) and the strongest signals (Figure V-3; Tables Sup V-6 and Sup V-7), although prec1 (143), long (113), and cld4 (108) were also predominant variables (Table V-1). The variable lat shared 68% of the SNPs with prec5 whereas long shared 36% of the SNPs with prec1 and 21% with cld4 (Table Sup V-10).

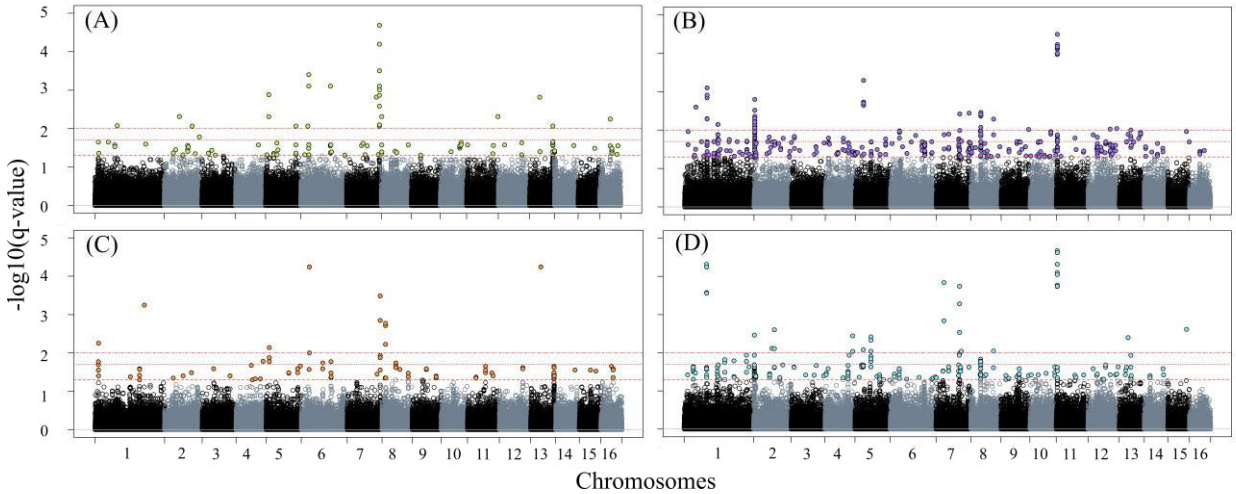


Figure V-3 - Manhattan plots representing the genome-wide distribution of significance values –log10 (*q*-value) obtained by LFMM for the environmental variables with the strongest associations. (A) prec1: 164 SNPs, (B) prec5: 596 SNPs, (C) long: 113 SNPs, (D) lat: 385 SNPs. The red lines indicate FDR values of 0.05, 0.02 and 0.01.

The SNPs detected by both GEA methods showed an enrichment in exons (P-value<2.20×10^{-16}), UTRs (P-value=0.018), introns <50 bp of exons (P-value=8.10×10^{-5}), and intergenic regions <2 Kb (P-value=1.75×10^{-12}) and <1 Kb upstream of exons (P-value=1.27×10^{-6}, $\chi 2$ test).

Table V-1 - Environmental variables and number of associated SNPs identified exclusively by LFMM or Samβada and simultaneously by both methods.

Environmental variables	LFMM	Samβada	Overlapping
Lat	35	64	350
Prec5	288	10	308
Prec1	21	105	143
Long	0	564	113
Cld4	89	23	108
Prec8	272	58	79
Tmin1	19	11	57
Ins4	48	28	20
Rh3	92	14	19
Rh1	17	3	16
Land cover	21	0	11
Cld7	26	3	8
Rh6	17	14	6
Tmin6	2	4	5

PCAdapt fast

A total of 107 outlier SNPs were identified by PCAdapt for an FDR<0.2 (Figure Sup V-5; Tables Sup V-9 and Sup V-11). Of the 107 SNPs, 26 were located in 10 intergenic regions and 81 in 22 genes. Interestingly, most genic SNPs (63%) marked only three genes: GB49881 (29), GB49882 (11), GB46620 (11). The remaining 30 SNPs were located in 19 genes, including the previously GEA-identified GB48105 (3 SNPs), GB54460 (2 SNPs), and GB40077 (1 SNP).

Putative targets of selection identified by PCAdapt fast were enriched in exons (P-value=1.60×10^{-5}), and in intergenic regions <2 Kb (P-value=5.00×10^{-3}) and <1Kb upstream (P-value=2.00×10^{-3}, χ^2 test) of exons.

The Strongest Candidate SNPs

A total of 830 SNPs was detected by at least two selection methods; 9.8% were located in exons (40 non-synonymous and 41 synonymous SNPs), 2.5% in UTRs (21 SNPs), 46.3% in introns (384 SNPs, of which 31 were <50 bp of exons), 16.3% in intergenic regions adjacent to (2-2,000 bp;

135 SNPs, of which 47 were <1 Kb upstream) exons and 25.2% distant from (2,001-145,825 bp; 209 SNPs) exons.

The 830 SNPs exhibited |iHs| values ranging from 0.005 to 7.2 (Tables Sup V-8 and Sup V-9). A total of 147 SNPs were strong candidates for recent ongoing selection as they showed a |iHs|>2 (Table Sup V-9). The two top-ranked SNPs displayed a |iHs|>7, standing out by a remarkably strong selection signature. One of these two is located 834 bp upstream of the undescribed gene GB54883 and the other is a longitude-associated non-synonymous SNP located in GB55263 (see further details in the protein modelling section).

The great majority (486 SNPs, 58.6%) of the 830 SNPs were located in exons, introns and UTRs of 181 genes. Of these 181 genes, 8 carried >10 SNPs (Tables Sup V-8 and Sup V-9), mostly associated with prec5 and lat (Table V-2). The aforementioned GB49881, GB49882, GB46620, and GB43005 are amongst the 8 genes and are highlighted by possessing 29, 19, 18 and 14 SNPs, respectively. Four genes were tagged by non-synonymous SNPs with GB48703 harbouring the most (Table V-2).

Table V-2 - Candidate genes containing more than 10 SNPs detected concurrently by at least two selection methods.

A. mellifera gene	# SNPs	SNPs distribution across genomic regions	Environmental variable
GB49881	29	29 Intronic	Long, prec1, cld4
GB48698	21	1 Exonic (syn), 17 intronic, 3 UTR	Lat, prec5
GB49882	19	5 Exonic (syn), 14 intronic	cld4
GB46620	18	2 Exonic (non-syn), 2 exonic (syn), 14 intronic	Lat, prec1, prec5, prec8, ins4
GB43005	14	1 Exonic (syn), 13 intronic	Lat, prec1, prec5, tmin1, tmin6, ins4
GB48703	13	4 Exonic (non-syn), 1 exonic (syn), 6 intronic, 2 UTR	Lat, prec5
GB48709	13	3 Exonic (non-syn), 1 exonic (syn), 9 intronic	Lat, prec5
GB48699	11	1 Exonic (non-syn), 10 intronic	Lat, prec5

Note.- Genes marked in bold carry SNPs that were cross-detected by iHS and/or PCAdapt.

The highest stringency cross-validation based on all methods (Samβada, LFMM and PCAdapt) identified 83 overlapping SNPs (Figure V-4 and Tables Sup V-S8 and Sup V-9). These were located in 14 genes, including GB49881 and GB46620 each containing >10 SNPs. From the 83 SNPs, 40 displayed elevated |iHs| values (>2.0) representing 4 genes and 5 intergenic regions (Table V-3). Genes with the highest number of SNPs and the uppermost |iHs| values were

GB49881 (28 SNPs, |iHs|>3.0) and GB49899 (4 SNPs, |iHs|>3.2). Interestingly, gene GB49881 is only 1,864 bp away from another strong candidate, GB49882. This remarkably short intergenic region contained 15 SNPs detected by the three methods with |iHs|>1.6 (Table Sup V-9).

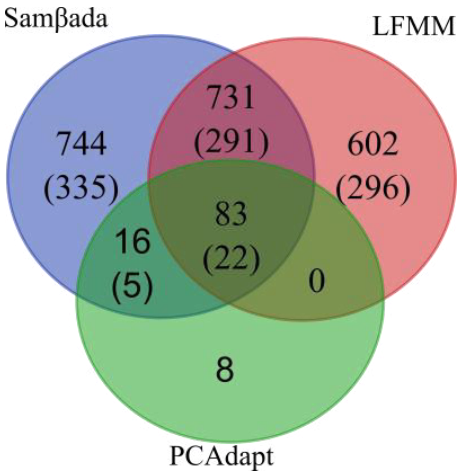


Figure V-4 - Overlapping SNPs identified by the three genome-scan methods based on different models and assumptions. Numbers in the intersection regions represent overlapping SNPs among two or three methods. Numbers in parentheses show the corresponding genes harbouring the SNPs.

The 830 SNPs showed an enrichment in exons (P-value<2.20×10^{-16}), UTRs (P-value=0.015), introns <50 bp of exons (P-value=4.32×10^{-5}), and intergenic regions <2 Kb (P-value=1.00×10^{-11}) and <1 Kb upstream of exons (P-value=2.58×10^{-6}, $\chi 2$ test).

Table V-3 - Genomic information, and associated environmental variables, of candidate genes cross-detected by Sam*β*ada, LFMM, PCAdapt and |iHs|>2.

A. mellifera **gene**	**# SNPs**	**Genomic position**	**Env. variables**	**Putative function**
GB49881	28	Intronic	Prec1, cld4	Undescribed
GB49899	4	Intronic	Long, prec1, cld4	Pdz domain
GB49874	2	Intergenic (<2060 bp)	Long, prec1, cld4	Undescribed
GB44109	1	Intergenic (2182 bp)	Prec1	Oxidation-reduction process
GB48105	1	Intronic (8 bp)	Long, prec1	Neurogenesis
GB49878	1	Intergenic (495 bp)	Long, prec1	Response to DDT
GB49879	1	Intronic (204 bp)	Long, prec1	Sleep
GB51286	1	Intergenic (17,753 bp)	Prec1	Undescribed
GB51427	1	Intergenic (952 bp)	Long	Humoral response

Note.- Genes detected by other honey bees studies (Harpur *et al.* 2014; Wallberg *et al.* 2014) are marked in bold. Orthologs in *D. melanogaster* were coded by FBgnxxxxxxx. Genes with no orthologs in other species were classified as orphans. Putative functions were summarized from FLYBASE.

Protein Modelling

A total of 40 non-synonymous SNPs was detected by at least two selection methods. The 40 SNPs were located in 26 genes. Protein prediction was available for only 10 of the 26 genes and 4 genes contained SNPs outside the 3D model (Table Sup V-12). The remaining 6 genes (GB40077, GB45499, GB47279, GB48707, GB51396, GB55263) were translated into a total of 34 protein variants (Table Sup V-13). Genes GB48707 and GB45499 were the least diverse, with 4 variants, and gene GB40077 was the most diverse, with 9 variants (Figure V-5, Figure Sup V-6, Table Sup V-13). Most of these protein variants exhibited lower energy minimization than the reference (11 variants) and values of $\Delta\Delta G>0$ (14 variants), with the highest $\Delta\Delta G$ values displayed by variants C and E of gene GB47279 (3.94 Kcal/mol and 2.29 Kcal/mol, respectively), indicating that the variants are less stable than the reference protein.

The geographical patterns of protein variation displayed by GB40077 (9 variants), GB45499 (4 variants), GB55263 (6 variants), GB47279 (6 variants), GB48707 (4 variants), and GB51396 (5 variants) are shown in Figure V-5 and Figure Sup V-6. While the two highly frequent protein variants (A and B) of genes GB45499 (55 A and 30 B, out of 87 individuals), GB55263 (56 A and 19 B, out of 87 individuals), and GB47279 (54 A and 28 B, out of 87 individuals) were oriented along a northeastern-southwestern axis, the more diverse GB40077 had a relatively high-frequency of variant B (16 out of 87 individuals), which was mostly restricted to the Atlantic side of Iberia.

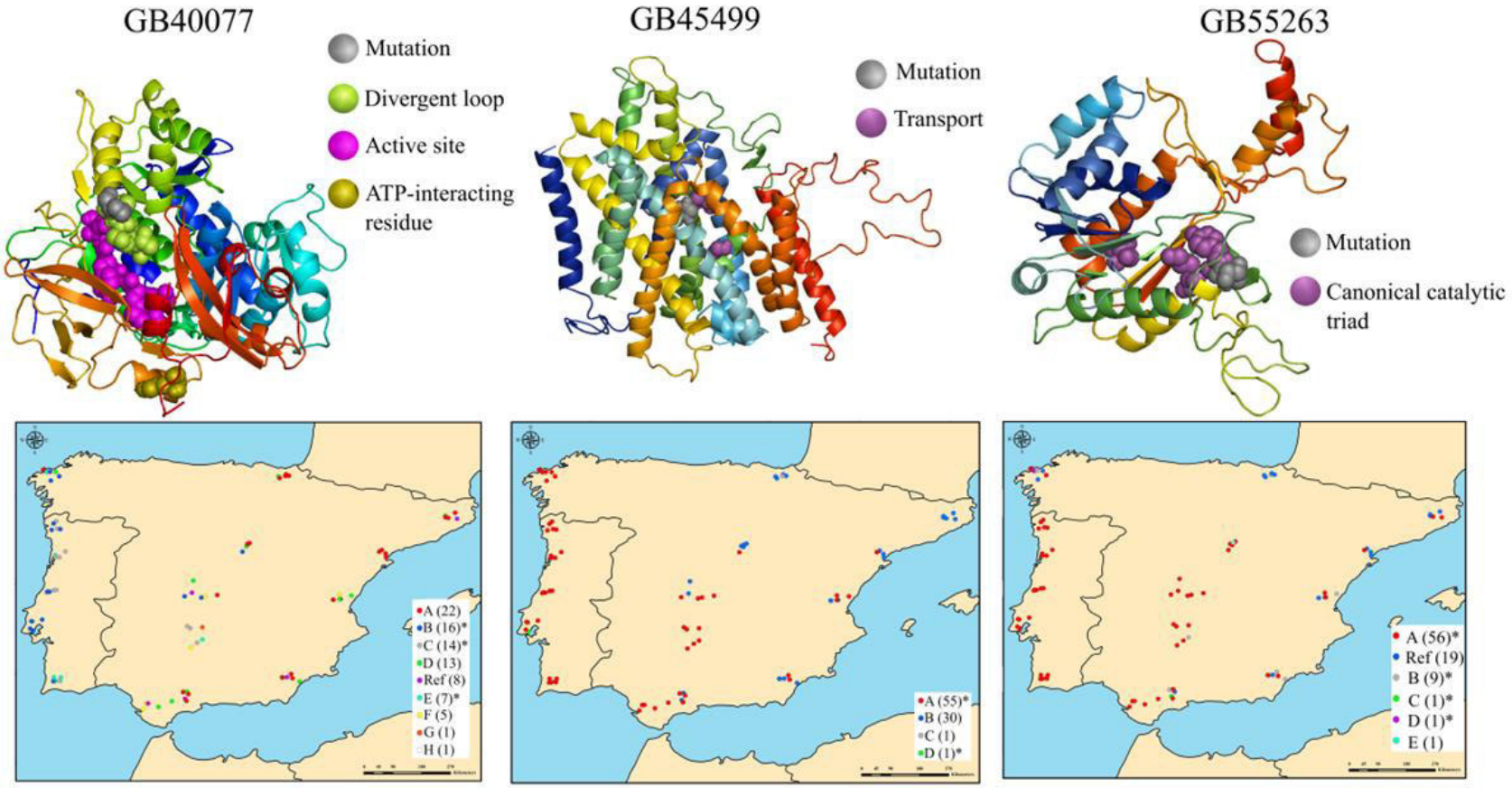


Figure V-5 - Predicted protein structures of the three genes harbouring non-synonymous candidate SNPs, detected by three genome-wide methods, located near to important places in the protein. The structures were predicted by Pymol considering the BeeBase reference amino acid sequences. The grey spheres represent the position and altered amino acids. The coloured spheres represent places with a known and important function in the protein. The asterisk represents the variants carrying the SNP under selection. The maps show how the different protein variants gather in space. Numbers in parentheses show the number of individual honey bees for each variant.

Gene Ontology and Annotation

The power of the gene ontology (GO) analysis for uncovering the biological significance of candidate regions identified in whole-genome selection scans depends on the number of annotated genes available for the focal organism (Yon Rhee *et al.*, 2008). Of the 181 candidate genes identified here by at least two selection methods, only 131 were retrieved from the DAVID database. Hence, the GO analysis should be interpreted with caution as it may reflect a biased representation of candidate genes and miss biological functions. The 131 genes showed a significant enrichment (P-value<0.05) for 8 functional terms (Table Sup V-15), of which 6 formed two clusters. The first cluster (enrichment score=1.96) included 4 terms related with *membrane* and the second (enrichment score=1.19) included 2 terms related with *immunoglobulin.* The remaining functional terms, *olfactory learning* and *positive regulation of canonical Wnt pathway,* were not clustered. While these 2 terms only included 2 genes each, the fold enrichment values were remarkably high (98.93 for *olfactory learning* and 65.95 for *positive regulation of canonical Wnt pathway*).

The *membrane* cluster comprised 31 genes of which 19 could be grouped into three classes of proteins, including cell-surface receptors (9), transport (7), and cell-adhesion (3; Tables V-14 and Sup V-15). The 9 cell-surface receptor genes belong to four families, including the G-protein-coupled receptor family (GB48005, GB49166, GB55784, GB40666, GB48703 and its paralogue GB48704), the ion-channel-coupled receptor (GB48639), the enzyme-coupled receptor (GB43446), and the CD36/scavenger receptor (GB49363). The 7 transport proteins were represented by potassium channels (GB49879 and its paralogue GB49882) and transporters (GB45499, GB50262, GB49320, GB46597, GB53142). Finally, the cell-adhesion proteins were represented by GB49778, GB53374, and GB44159. The *immunoglobulin* cluster was represented by 5 genes, including GB40061, GB40069, and GB48691, which are related with sensory perception of chemical stimulus (the latter more specifically related with olfaction), and GB49778 and GB44159, which are related with cell adhesion (Tables Sup V-8 and Sup V-15).

Although unrepresented in the GO enrichment analysis, many other genes are good candidates for local adaptation in the IHB as they are putatively implicated in the same biological function. These functions include reproduction with 14 genes, immunity with 12 genes, regulation of transcription with 11 genes, lipid storage and biosynthesis with 10 genes, olfaction with 9 genes, vision with 5 genes, and detoxification with 3 genes (Table Sup V-14).

Discussion

In this study we employed conceptually different analytical tools to disentangle signatures of selection from genome-wide geospatial variation in the IHB. By scanning 87 whole genomes, we were able to refine inferences previously made from a limited number of pre-ascertained biased SNPs (Chávez-Galarza *et al.*, 2013) and provide unique insights into the molecular basis of local adaption in the IHB. In addition to obtaining unbiased information about the type of genes and biological processes putatively underlying local adaptation, we have never been so close to finding associated causal mutations.

The majority of the 40 non-synonymous outlier SNPs are likely causal mutations, especially those laying in genes GB40077, GB55263, and GB45499, as they could be linked to amino acid positions important for protein functioning (Campos *et al.*, 2013; Chhabra *et al.*, 2012; Hu, 2010; Watanabe *et al.*, 2010). On the other hand, it is possible that many outlier SNPs are hitchhiked with the actual genetic target of the selective event as 66% of the outlier SNPs were <5 Kb apart. Yet, the hypothesis that many linked multiple causal mutations have a functional role cannot be ruled out (Anderson, 2010; Lamichhaney *et al.*, 2015; Martinez Barrio *et al.*, 2016). Further identification of causal mutations is a challenging endeavour that will require more accurate and comprehensive annotations of the honey bee genome, and especially annotation of the non-coding regulatory DNA, along with evidence from biochemical and functional assays.

Gene regulation is the primary source of adaptive change in the Iberian honey bee

A great proportion (90.2%) of cross-detected outlier SNPs are located in non-coding DNA, as opposed to the 9.8% exonic SNPs. A similar disproportionate fraction of non-coding to coding SNPs has been identified by whole-genome scans in other organisms, including humans (Bertrand *et al.*, 2015; Vernot *et al.*, 2012), fishes (Bertrand *et al.*, 2015; Martinez Barrio *et al.*, 2016), and fruit flies (Andolfatto, 2005; Božičević *et al.*, 2016). The majority of outlier SNPs beyond the exome suggests that mutations in regulatory sequences, rather than mutations in protein-coding sequences, are the primary source of adaptive change in the IHB.

Non-coding SNP variants have been implicated in almost all processes of gene regulation, ranging from transcription to post-translation (Li *et al.*, 2015). Significant enrichment of outlier SNPs laying within 1 Kb upstream from the transcription start site of 23 genes (e.g. GB51427,

GB48700, GB48697), where the promoter is expected to be located, strongly suggests that *cis*-regulatory changes underlie local adaptation in IHBs. Intronic outlier SNPs in the immediate vicinity of exons (<50 bp) were also found to be enriched. It is possible that some outlier intronic SNPs are involved in regulating gene expression through, for example, alternative splicing. The strong and adjacent (1,864 bp apart) candidate genes GB49881 and GB49882, each containing 29 and 14 intronic outlier SNPs, respectively, provide an interesting case potentially involving this mechanism. These genes share transcript sequence associated with the SHAW protein, a member of voltage-gated K^+ channels, for which there is five alternative variants described. The remarkable number of outlier SNPs identified in GB49881 and GB49882 strongly suggests that this level of variation is important for regulating gene expression possibly by alternative splicing determining when and where each variant is translated (Ast, 2004). Alternative splicing has recently been implicated in local adaptation involving circadian clock genes (Kaiser *et al.*, 2016).

The other parts of the genome known to be involved in regulation are UTRs (Barrett *et al.*, 2012). Because most nucleotides in UTRs are functionally important (Andolfatto, 2005), it is possible that most, if not all, 21 SNPs in the UTR-enriched category are the direct targets of selection. In addition to *cis*-regulatory mechanisms, the 11 candidate genes related with transcription regulation and the GO significant enrichment in the term *positive regulation of canonical Wnt pathway* strongly suggest that *trans* regulation is implicated in IHB local adaptation.

Our findings add to a growing body of evidence supporting the idea that local adaptation is not primarily driven by protein-coding sequences, as previously thought, but rather regulatory mutations might even play a disproportionate part in the evolution of quantitative traits and of responses to stressors, resources and pathogens (reviewed by Wray, 2007).

Candidate Genes for Local Adaptation

Support for selection is provided by functional annotations of candidate genes that can be directly related to colony fitness and it is particularly compelling when multiple candidate genes are implicated in the same biological function. While the GO enrichment analysis only detected 8 significant terms (4 related with *membrane*), functional annotations indicate that many fitness-related functions are represented by multiple genes (e. g. as reproduction with 14 genes or immunity with 12 genes). Other fitness-related biological functions were highlighted as they displayed strong selection signals. These include olfaction, circadian clock, and lipids biosynthesis

and storage. Many of the candidate genes identified for the IHB were also detected by whole-genome selection scans for other honey bee subspecies (Table Sup V - 8; Fuller *et al.*, 2015; Harpur *et al.*, 2014; Wallberg *et al.*, 2014), suggesting that they are adaptively important across diverse environments.

Membrane proteins. GO analysis identified a cluster of 31 candidate genes encoding for membrane proteins in the IHB. The importance of membrane proteins in adaptation to new environments is evidenced by their rapid evolution compared with cytosolic proteins (Sojo *et al.*, 2016). In this study, three candidate genes are highlighted in the group of membrane transport proteins. Gene GB45499 is one of strongest candidates for selection as it carries a mutation leading to an amino acid change in a site of the protein located in the transmembrane region and involved in transport activity (Watanabe *et al.*, 2010). The replaced amino acid is located in the alpha-helix and, together with two other amino acids, it is important to maintain the open pathway from the intracellular space (Watanabe *et al.*, 2010). Genes GB49879 and GB49882 are paralogous encoding voltage-gated K^+ channels, which are putatively implicated in the circadian clock (see below). GB49879 was detected by all selection methods and exhibits a |iHS|=2.19, indicative of strong signals of ongoing selection (Voight *et al.*, 2006). GB49882 is tagged by 19 SNPs of which five are exonic. In addition to membrane transport protein, the selection scan identified four candidate genes in the group of membrane receptors all implicated in olfaction.

Olfaction. The adaptive relevance of olfaction is revealed by the significant enrichment of the GO term *olfactory learning* and identification of nine candidate genes, including the membrane receptors. GB48703 and GB48691 are amongst the most striking candidates deserving further investigation. GB48703 encodes an olfactory membrane receptor and carries 13 outlier SNPs, of which four are non-synonymous. GB48691 is implicated in olfactory learning and carries nine outlier SNPs, of which one is non-synonymous. Unfortunately, protein prediction was not available for these genes hampering inferences on effects of the non-synonymous mutations in protein functioning. Colony fitness relies largely on olfactory perception. Olfaction is implicated in the learning process, which is crucial for the improvement of resources' acquisition, as well as in a wide array of behaviours, including detection of possible dangers, recognition of potential mates, and social interactions (Sandoz, 2011; Schowalter, 2016). Olfaction has also been shown to play a

major role in the detection of brood cells infested by *Varroa destructor* (Navajas *et al.*, 2008), an invasive mite that has been challenging honey bee health at unprecedented levels.

Circadian clock. The honey bee relies on a circadian clock to synchronize foraging behaviour and reproductive swarming with the maximum daily and seasonal availability of food resources (Bloch, 2010; Simpson, 1958). The importance of circadian rhythmicity in local adaption of IHBs is suggested by five candidate genes putatively operating in the three functional components of the circadian clock. The clock component "input pathways" is represented by GB48005. Its putative orthologue in *Drosophila* encodes for a G protein-coupled serotonin receptor (Gasque *et al.*, 2013), which regulates the entrainment of circadian behavioural rhythms affecting the molecular response to light (Yuan *et al.*, 2005). The core component "oscillator" is represented by GB52077. Its putative orthologue in *Drosophila* encodes for the transcription factor Period (*Per*). The honey bee *amPer* is an essential element of circadian rhythmicity, and its product is involved in a negative transcription/translational auto-regulatory feedback loop (Eban-Rothschild & Bloch, 2012). The development of strong circadian rhythms in honey bee foragers has been shown to be associated with changes in brain *Per* expression (Bloch, 2010). Finally, the component "output pathways" is represented by the striking candidates GB49879, GB49881 and GB49882. Genes GB49881 and GB49882 share transcript sequence associated with the SHAW voltage-gated K^+ channel protein. GB49881 and GB49879 are paralogous encoding for SHAW and SHAW-like proteins, respectively. In *Drosophila,* the SHAW potassium channels regulate the intrinsic excitability in all neurons, being therefore important for output rhythms of the circadian clock (Hodge & Stanewsky, 2008). The five clock genes were mostly marked by intronic outlier SNPs, suggesting that gene regulation might be the predominant molecular mechanism to meet functional demands of circadian rhythmicity.

Lipid biosynthesis and storage. Ten lipid-related candidate genes mostly related with lipids biosynthesis and storage were detected in the IHB and GB55263 and GB40077 are amongst the top-ranked candidates possibly playing a central role in IHB adaptation. The non-synonymous SNPs mapped to these genes are likely causal mutations as they lead to replacement of amino acids located in functionally important sites of the proteins. The mutation in GB55263 leads to an amino acid replacement in a canonical catalytic triad (Chhabra *et al.*, 2012). The mutation in GB40077 leads to an amino acid replacement in a divergent loop (Hu, 2010), which is important for mediating the protein-protein interaction or is part of the ATP binding site (Campos *et al.*, 2013).

The *Drosophila* ortholog of GB40077 is implicated in lipid homeostasis (Xu *et al.*, 2012) and has been linked to the circadian clock (Claridge-Chang *et al.*, 2001; Xu *et al.*, 2011).

Driving Forces of Local Adaptation

Precipitation and cloud cover. Precipitation (in May and January) and cloud cover (in April) are the variables most frequently and strongly associated with SNPs. While precipitation and cloud cover may act as selective pressures by interfering with foraging, winter mortality, behaviour in the nest, mating flights (Bol'shakova, 1978; Lensky & Demter, 1985), whether they are direct causes of selection is unclear. It may very well be that these climatic factors operate indirectly by determining availability of pollen and nectar sources across space and time which will not only influence foraging and colony build up but also reproduction. Due to the highly contrasting climates (e.g. average annual precipitation is 1336.3 mm in the northwest and 284.6 mm in the southeast), plant communities (wild plants or crops) and blooming seasons are very heterogeneous across Iberia (Loidi, 2017). This could favour evolution of locally adapted populations to food resources. An interesting example of such adaptation is provided by the existence of an ecotype of *A. m. mellifera* (the other M-lineage subspecies in Europe) that has an annual brood cycle fine-tuned with the phenology of an abundant floral source in the Landes region in France (Strange *et al.*, 2007; Strange *et al.*, 2008).

Precipitation in January and May and cloud cover in April covary with temperature, insolation, cloud cover, and precipitation in other months. Multicollinearity may lead to incorrectly identifying a variable as causal when the true selective pressure is a correlated variable. However, it is also possible that selection is driven by composite environmental cues. For example, the mating behaviour of honey bee queens is influenced by a combination of temperature, wind, and cloud cover (Bol'shakova, 1978; Lensky & Demter, 1985).

Latitude and longitude. Latitude, and to a lesser extent longitude, showed a large number of associations with outlier SNPs (677 and 449, respectively). While latitude and longitude do not act directly on organisms, they serve as composite variables representing multiple environmental factors, anyone or a combination of which could be exerting parallel selective forces. Latitude has been found to be associated with circadian clock genes in *Drosophila* (Costa *et al.*, 1992; Kyriacou *et al.*, 2007; Tauber *et al.*, 2007; Yerushalmi & Green, 2009) and humans (Dall'Ara *et al.*, 2016; Forni *et al.*, 2014), and now in honey bees. Clock genes are tagged by SNPs forming latitudinal

and longitudinal gradients in Iberia. This finding suggests that circadian rhythmicity is involved in local adaptation in IHBs by matching important behaviours, such as feeding and reproduction, with the diverse daily and seasonal environmental oscillations of Iberia.

Insights into genetic adaptation to the broad range of Iberian environments.

Using both genetic and environmental data we identified candidate genes putatively under climate-driven adaptation. This information is particularly important in the context of rapid global climate change, helping us to understand the mechanisms employed by organisms to adapt to varying environmental conditions. The circadian clock allows an organism to anticipate and adapt to predictable environmental changes. Because the environmental conditions are variable, clock systems have some plasticity (Yerushalmi & Green, 2009); however, with the rapid global changes characterized by unpredictability, the circadian and seasonal rhythms could be impaired and phenology changes are described as one of the first effects of global climate change (Helm *et al.*, 2013; Peñuelas & Filella, 2001).

Acknowledgments

This work was supported by Fundação para a Ciência e Tecnologia (FCT) through the programs COMPETE/QREN/EU (PTDC/BIA-BEC/099640/2008) and the 2013-2014 BiodivERsA/FACCE-JPI (joint call for research proposals, with the national funders FCT, Portugal, CNRS, France, and MEC, Spain) to MAP, and the PhD scholarship SFRH/BD/84195/2012 to DH. John C. Patton, Phillip San Miguel, Paul Parker, Rick Westerman, University of Purdue, sequenced the 87 whole genomes of IHBs. José Rufino provided computational resources at IPB. Analyses were performed using the computational resources at the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX), Uppsala University.

References

Anderson, E. C. (2010). Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. *Mol Ecol Resour, 10*(4), 701-710. doi: 10.1111/j.1755-0998.2010.02846.x

Andolfatto, P. (2005). Adaptive evolution of non-coding DNA in Drosophila. *Nature, 437*(7062), 1149-1152. doi: 10.1038/nature04107

Ast, G. (2004). How did alternative splicing evolve? *Nature Reviews Genetics, 5*(10), 773-782. doi: 10.1038/nrg1451

Barrett, L. W., Fletcher, S., & Wilton, S. D. (2012). Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cellular and Molecular Life Sciences, 69*(21), 3613-3634. doi: 10.1007/s00018-012-0990-9

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological, 57*(1), 289-300.

Benjelloun, B., Alberto, F. J., Streeter, I., Boyer, F., Coissac, E., Stucki, S., BenBati, M., Ibnelbachyr, M., Chentouf, M., Bechchari, A., Leempoel, K., Alberti, A., Engelen, S., Chikhi, A., Clarke, L., Flicek, P., Joost, S., Taberlet, P., Pompanon, F., & NextGen, C. (2015). Characterizing neutral genomic diversity and selection signatures in indigenous populations of Moroccan goats (*Capra hircus*) using WGS data. *Front Genet, 6*, 107. doi: 10.3389/fgene.2015.00107

Bertrand, B., Alburaki, M., Legout, H., Moulin, S., Mougel, F., & Garnery, L. (2015). MtDNA COI-COII marker and drone congregation area: an efficient method to establish and monitor honeybee (*Apis mellifera L.*) conservation centres. *Mol Ecol Resour, 15*(3), 673-683. doi: 10.1111/1755-0998.12339

Beye, M., Gattermeier, I., Hasselmann, M., Gempe, T., Schioett, M., Baines, J. F., Schlipalius, D., Mougel, F., Emore, C., Rueppell, O., Sirvio, A., Guzmán-Novoa, E., Hunt, G., Solignac, M., & Page, R. E., Jr. (2006). Exceptionally high levels of recombination across the honey bee genome. *Genome Res, 16*(11), 1339-1344. doi: 10.1101/gr.5680406

Bhattacharya, D., Nowotny, J., Cao, R., & Cheng, J. (2016). 3Drefine: an interactive web server for efficient protein structure refinement. *Nucleic Acids Res, 44*(W1), W406-W409. doi: 10.1093/nar/gkw336

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Gallo Cassarino, T., Bertoni, M., Bordoli, L., & Schwede, T. (2014). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res, 42*(Web Server issue), W252-258. doi: 10.1093/nar/gku340

Biswas, S., & Akey, J. M. (2006). Genomic insights into positive selection. *Trends in Genetics, 22*(8), 437-446. doi: http://dx.doi.org/10.1016/j.tig.2006.06.005

Bloch, G. (2010). The social clock of the honeybee. *J Biol Rhythms, 25*(5), 307-317. doi: doi:10.1177/0748730410380149

Bol'shakova, M. D. (1978). The flight of honey bee drones, *Apis mellifera L.* (Hymenoptera, Apidae), to the queen in relation to various ecological factors. *Entomological Review, 56*, 53-56.

Božičević, V., Hutter, S., Stephan, W., & Wollstein, A. (2016). Population genetic evidence for cold adaptation in European *Drosophila melanogaster* populations. *Mol Ecol, 25*(5), 1175-1191. doi: 10.1111/mec.13464

Campana, M. G., Hunt, H. V., Jones, H., & White, J. (2011). CorrSieve: software for summarizing and evaluating Structure output. *Mol Ecol Resour, 11*(2), 349-352. doi: 10.1111/j.1755-0998.2010.02917.x

Campos, B. M., Sforca, M. L., Ambrosio, A. L., Domingues, M. N., Brasil de Souza Tde, A., Barbosa, J. A., Paes Leme, A. F., Perez, C. A., Whittaker, S. B., Murakami, M. T., Zeri, A. C., & Benedetti, C. E. (2013). A redox 2-Cys mechanism regulates the catalytic activity of divergent cyclophilins. *Plant Physiol, 162*(3), 1311-1323. doi: 10.1104/pp.113.218339

Cánovas, F., Rúa, P., Serrano, J., & Galián, J. (2011). Microsatellite variability reveals beekeeping influences on Iberian honeybee populations. *Apidologie, 42*(3), 235-251. doi: 10.1007/s13592-011-0020-1

Chávez-Galarza, J., Henriques, D., Johnston, J. S., Azevedo, J. C., Patton, J. C., Munoz, I., De la Rua, P., & Pinto, M. A. (2013). Signatures of selection in the Iberian honey bee (*Apis mellifera iberiensis*) revealed by a genome scan analysis of single nucleotide polymorphisms. *Mol Ecol, 22*(23), 5890-5907. doi: 10.1111/mec.12537

Chávez-Galarza, J., Henriques, D., Johnston, J. S., Carneiro, M., Rufino, J., Patton, J. C., & Pinto, M. A. (2015). Revisiting the Iberian honey bee (*Apis mellifera iberiensis*) contact zone: maternal and genome-wide nuclear variations provide support for secondary contact from historical refugia. *Mol Ecol, 24*(12), 2973-2992. doi: 10.1111/mec.13223

Chen, C., Liu, Z., Pan, Q., Chen, X., Wang, H., Guo, H., Liu, S., Lu, H., Tian, S., Li, R., & Shi, W. (2016). Genomic analyses reveal demographic history and temperate adaptation of the newly discovered honey bee subspecies *Apis mellifera sinisxinyuan n. ssp*. *Mol Biol Evol, 33*(5), 1337-1348. doi: 10.1093/molbev/msw017

Chhabra, A., Haque, A. S., Pal, R. K., Goyal, A., Rai, R., Joshi, S., Panjikar, S., Pasha, S., Sankaranarayanan, R., & Gokhale, R. S. (2012). Nonprocessive [2 + 2]e- off-loading reductase domains from mycobacterial nonribosomal peptide synthetases. *Proc Natl Acad Sci U S A, 109*(15), 5681-5686. doi: 10.1073/pnas.1118680109

Claridge-Chang, A., Wijnen, H., Naef, F., Boothroyd, C., Rajewsky, N., & Young, M. W. (2001). Circadian regulation of gene expression systems in the Drosophila head. *Neuron, 32*(4), 657-671.

Coop, G., Witonsky, D., Di Rienzo, A., & Pritchard, J. K. (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics, 185*(4), 1411-1423.

Costa, R., Peixoto, A. A., Barbujani, G., & Kyriacou, C. P. (1992). A latitudinal cline in a Drosophila clock gene. *Proc Biol Sci, 250*(1327), 43-49. doi: 10.1098/rspb.1992.0128

Dall'Ara, I., Ghirotto, S., Ingusci, S., Bagarolo, G., Bertolucci, C., & Barbujani, G. (2016). Demographic history and adaptation account for clock gene diversity in humans. *Heredity (Edinb), 117*(3), 165-172. doi: 10.1038/hdy.2016.39

de Villemereuil, P., Frichot, E., Bazin, E., Francois, O., & Gaggiotti, O. E. (2014). Genome scan methods against more complex models: when and how much should we trust them? *Mol Ecol, 23*(8), 2006-2019. doi: 10.1111/mec.12705

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet, 43*(5), 491-498. doi: 10.1038/ng.806

Duforet-Frebourg, N., Bazin, E., & Blum, M. G. (2014). Genome scans for detecting footprints of local adaptation using a Bayesian factor model. *Mol Biol Evol, 31*(9), 2483-2495. doi: 10.1093/molbev/msu182

Duforet-Frebourg, N., Luu, K., Laval, G., Bazin, E., & Blum, M. G. (2015). Detecting genomic signatures of natural selection with Principal Component Analysis: Application to the 1000 genomes data. *Mol Biol Evol*. doi: 10.1093/molbev/msv334

Eban-Rothschild, A., & Bloch, G. (2012). Circadian rhythms and sleep in honey bees *Honeybee neurobiology and behavior* (pp. 31-45): Springer.

Engel, M., S. (1998). Fossil honey bees and evolution in the genus Apis (Hymenoptera: Apidae). *Apidologie, 29*(3), 265-281.

Excoffier, L., Hofer, T., & Foll, M. (2009). Detecting loci under selection in a hierarchically structured population. *Heredity (Edinb), 103*(4), 285-298. doi: 10.1038/hdy.2009.74

Foll, M., & Gaggiotti, O. (2008). A Genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics, 180*(2), 977-993.

Forester, B. R., Jones, M. R., Joost, S., Landguth, E. L., & Lasky, J. R. (2016). Detecting spatial genetic signatures of local adaptation in heterogeneous landscapes. *Mol Ecol, 25*(1), 104-120. doi: 10.1111/mec.13476

Forni, D., Pozzoli, U., Cagliani, R., Tresoldi, C., Menozzi, G., Riva, S., Guerini, F. R., Comi, G. P., Bolognesi, E., Bresolin, N., Clerici, M., & Sironi, M. (2014). Genetic adaptation of the human circadian clock to day-length latitudinal variations and relevance for affective disorders. *Genome Biol, 15*(10), 499. doi: 10.1186/s13059-014-0499-7

Franck, P., Garnery, L., Solignac, M., & Cornuet, J. M. (1998). The origin of west European subspecies of honeybees (*Apis mellifera*): New insights from microsatellite and mitochondrial data. *Evolution, 52*(4), 1119-1134.

Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., & Francois, O. (2014). Fast and efficient estimation of individual ancestry coefficients. *Genetics, 196*(4), 973-983. doi: 10.1534/genetics.113.160572

Frichot, E., Schoville, S. D., Bouchard, G., & Francois, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol Biol Evol, 30*(7), 1687-1699. doi: 10.1093/molbev/mst063

Frichot, E., Schoville, S. D., de Villemereuil, P., Gaggiotti, O. E., & Francois, O. (2015). Detecting adaptive evolution based on association with ecological gradients: orientation matters! *Heredity (Edinb), 115*(1), 22-28. doi: 10.1038/hdy.2015.7

Fuller, Z. L., Nino, E. L., Patch, H. M., Bedoya-Reina, O. C., Baumgarten, T., Muli, E., Mumoki, F., Ratan, A., McGraw, J., Frazier, M., Masiga, D., Schuster, S., Grozinger, C. M., & Miller, W. (2015). Genome-wide analysis of signatures of selection in populations of African honey bees (*Apis mellifera*) using new web-based tools. *BMC Genomics, 16*, 518. doi: 10.1186/s12864-015-1712-0

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907 [q-bio.GN]*.

Garud, N. R., Messer, P. W., Buzbas, E. O., & Petrov, D. A. (2015). Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genet, 11*(2), e1005004. doi: 10.1371/journal.pgen.1005004

Gasque, G., Conway, S., Huang, J., Rao, Y., & Vosshall, L. B. (2013). Small molecule drug screening in Drosophila identifies the 5HT2A receptor as a feeding modulation target. *Scientific Reports, 3*, srep02120. doi: 10.1038/srep02120

Guex, N., & Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis, 18*(15), 2714-2723. doi: 10.1002/elps.1150181505

Guillot, G., Vitalis, R., le Rouzic, A., & Gautier, M. (2014). Detecting correlation between allele frequencies and environmental variables as a signature of selection. A fast computational approach for genome-wide studies. *Spatial Statistics, 8*, 145-155.

Harpur, B. A., Kent, C. F., Molodtsova, D., Lebon, J. M., Alqarni, A. S., Owayss, A. A., & Zayed, A. (2014). Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. *Proc Natl Acad Sci U S A, 111*(7), 2614-2619. doi: 10.1073/pnas.1315506111

Helm, B., Ben-Shlomo, R., Sheriff, M. J., Hut, R. A., Foster, R., Barnes, B. M., & Dominoni, D. (2013). Annual rhythms that underlie phenology: biological time-keeping meets environmental change. *Proceedings of the Royal Society B: Biological Sciences, 280*(1765). doi: 10.1098/rspb.2013.0016

Heymann, Y., Steenmans, C., Croissile, G., & Bossard, M. (1994). *Corine Land Cover*. Luxembourg.

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., Poss, M. L., Reed, L. K., Storfer, A., & Whitlock, M. C. (2016). Finding the genomic basis of local adaptation: Pitfalls, practical solutions, and future directions. *Am Nat, 188*(4), 379-397. doi: 10.1086/688018

Hodge, J. J., & Stanewsky, R. (2008). Function of the Shaw potassium channel within the Drosophila circadian clock. *PLoS One, 3*(5), e2274. doi: 10.1371/journal.pone.0002274

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet, 5*(6), e1000529. doi: 10.1371/journal.pgen.1000529

Hu, Y. (2010). Crystal structures and enzymatic mechanisms of a *Populus tomentosa* 4-coumarate–CoA ligase. doi: 10.2210/pdb3a9u/pdb

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res, 37*(1), 1-13. doi: 10.1093/nar/gkn923

Jensen, J. D., Kim, Y., DuMont, V. B., Aquadro, C. F., & Bustamante, C. D. (2005). Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics, 170*(3), 1401-1410. doi: 10.1534/genetics.104.038224

Joost, S., Bonin, A., Bruford, M. W., Despres, L., Conord, C., Erhardt, G., & Taberlet, P. (2007). A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Mol Ecol, 16*(18), 3955-3969. doi: 10.1111/j.1365-294X.2007.03442.x

Kaiser, T. S., Poehn, B., Szkiba, D., Preussner, M., Sedlazeck, F. J., Zrim, A., Neumann, T., Nguyen, L.-T., Betancourt, A. J., Hummel, T., Vogel, H., Dorner, S., Heyd, F., von Haeseler, A., & Tessmar-Raible, K. (2016). The genomic basis of circadian and circalunar timing adaptations in a midge. *Nature, 540*(7631), 69-73. doi: 10.1038/nature20151

Kang, L., Aggarwal, D. D., Rashkovetsky, E., Korol, A. B., & Michalak, P. (2016). Rapid genomic changes in *Drosophila melanogaster* adapting to desiccation stress in an experimental evolution system. *BMC Genomics, 17*, 233. doi: 10.1186/s12864-016-2556-y

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc, 10*(6), 845-858. doi: 10.1038/nprot.2015.053

Ko, A., Cantor, R. M., Weissglas-Volkov, D., Nikkola, E., Reddy, P. M., Sinsheimer, J. S., Pasaniuc, B., Brown, R., Alvarez, M., Rodriguez, A., Rodriguez-Guillen, R., Bautista, I. C., Arellano-Campos, O., Munoz-Hernandez, L. L., Salomaa, V., Kaprio, J., Jula, A., Jauhiainen, M., Heliovaara, M., Raitakari, O., Lehtimaki, T., Eriksson, J. G., Perola, M., Lohmueller, K. E., Matikainen, N., Taskinen, M. R., Rodriguez-Torres, M., Riba, L., Tusie-Luna, T., Aguilar-Salinas, C. A., & Pajukanta, P. (2014).

Amerindian-specific regions under positive selection harbour new lipid variants in Latinos. *Nat Commun, 5*, 3983. doi: 10.1038/ncomms4983

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., & Mayrose, I. (2015). Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Mol Ecol Resour, 15*(5), 1179-1191. doi: 10.1111/1755-0998.12387

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol, 33*(7), 1870-1874. doi: 10.1093/molbev/msw054

Kyriacou, C. P., Peixoto, A. A., Sandrelli, F., Costa, R., & Tauber, E. (2007). Clines in clock genes: fine-tuning circadian rhythms to the environment. *Trends in Genetics, 24*(3), 124-132. doi: 10.1016/j.tig.2007.12.003

Lai, F. N., Zhai, H. L., Cheng, M., Ma, J. Y., Cheng, S. F., Ge, W., Zhang, G. L., Wang, J. J., Zhang, R. Q., Wang, X., Min, L. J., Song, J. Z., & Shen, W. (2016). Whole-genome scanning for the litter size trait associated genes and SNPs under selection in dairy goat (*Capra hircus*). *Scientific Reports, 6*, 12. doi: 10.1038/srep38096

Lamichhaney, S., Berglund, J., Almen, M. S., Maqbool, K., Grabherr, M., Martinez-Barrio, A., Promerova, M., Rubin, C.-J., Wang, C., Zamani, N., Grant, B. R., Grant, P. R., Webster, M. T., & Andersson, L. (2015). Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature, 518*(7539), 371-375. doi: 10.1038/nature14181

Lensky, Y., & Demter, M. (1985). Mating flights of the queen honeybee (*Apis mellifera*) in a subtropical climate. *Comparative Biochemistry and Physiology Part A: Physiology, 81*(2), 229-241. doi: http://dx.doi.org/10.1016/0300-9629(85)90127-6

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics, 26*(5), 589-595. doi: 10.1093/bioinformatics/btp698

Li, M. J., Yan, B., Sham, P. C., & Wang, J. (2015). Exploring the function of genetic variants in the non-coding genomic regions: approaches for identifying human regulatory variants affecting gene expression. *Briefings in Bioinformatics, 16*(3), 393-412. doi: 10.1093/bib/bbu018

Loidi, J. (2017). *The vegetation of the Iberian Peninsula, Volume 2* (Vol. 2): Springer.

Lotterhos, K. E., & Whitlock, M. C. (2015). The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Mol Ecol, 24*(5), 1031-1046. doi: 10.1111/mec.13100

Luikart, G., England, P. R., Tallmon, D., Jordan, S., & Taberlet, P. (2003). The power and promise of population genomics: from genotyping to genome typing. *Nat Rev Genet, 4*(12), 981-994.

Lv, F. H., Agha, S., Kantanen, J., Colli, L., Stucki, S., Kijas, J. W., Joost, S., Li, M. H., & Ajmone Marsan, P. (2014). Adaptations to climate-mediated selective pressures in sheep. *Mol Biol Evol, 31*(12), 3324-3343. doi: 10.1093/molbev/msu264

MacCallum, C., & Hill, E. (2006). Being Positive about Selection. *PLoS Biol, 4*(3), e87. doi: 10.1371/journal.pbio.0040087

Makinen, H., Vasemagi, A., McGinnity, P., Cross, T. F., & Primmer, C. R. (2015). Population genomic analyses of early-phase Atlantic Salmon (*Salmo salar*) domestication/captive breeding. *Evol Appl, 8*(1), 93-107. doi: 10.1111/eva.12230

Manel, S., Joost, S., Epperson, B. K., Holderegger, R., Storfer, A., Rosenberg, M. S., Scribner, K. T., Bonin, A., & Fortin, M. J. (2010). Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field. *Mol Ecol, 19*(17), 3760-3772. doi: 10.1111/j.1365-294X.2010.04717.x

Manel, S., Perrier, C., Pratlong, M., Abi-Rached, L., Paganini, J., Pontarotti, P., & Aurelle, D. (2016). Genomic resources and their influence on the detection of the signal of positive selection in genome scans. *Mol Ecol, 25*(1), 170-184. doi: 10.1111/mec.13468

Martinez Barrio, A., Lamichhaney, S., Fan, G., Rafati, N., Pettersson, M., Zhang, H., Dainat, J., Ekman, D., Höppner, M., Jern, P., Martin, M., Nystedt, B., Liu, X., Chen, W., Liang, X., Shi, C., Fu, Y., Ma, K., Zhan, X., Feng, C., Gustafson, U., Rubin, C.-J., Sällman Almén, M., Blass, M., Casini, M., Folkvord, A., Laikre, L., Ryman, N., Ming-Yuen Lee, S., Xu, X., & Andersson, L. (2016). The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *eLife, 5*, e12081. doi: 10.7554/eLife.12081

Meirmans, P. G. (2012). The trouble with isolation by distance. *Mol Ecol, 21*(12), 2839-2846. doi: 10.1111/j.1365-294X.2012.05578.x

Meixner, M. D., Leta, M. A., Koeniger, N., & Fuchs, S. (2011). The honey bees of Ethiopia represent a new subspecies of *Apis mellifera- Apis mellifera simensis* n. ssp. . *Apidologie, 42*, 425-437.

Miguel, I., Baylac, M., Iriondo, M., Manzano, C., Garnery, L., & Estonba, A. (2011). Both geometric morphometric and microsatellite data consistently support the differentiation of the *Apis mellifera* M evolutionary branch. *Apidologie, 42*(2), 150-161. doi: 10.1051/apido/2010048

Miguel, I., Iriondo, M., Garnery, L., Sheppard, W. S., & Estonba, A. (2007). Gene flow within the M evolutionary lineage of *Apis mellifera*: role of the Pyrenees, isolation by distance and post-glacial re-colonization routes in the western Europe. *Apidologie, 38*(2), 141-155. doi: 10.1051/apido:2007007

Mikheyev, A. S., Tin, M. M., Arora, J., & Seeley, T. D. (2015). Museum samples reveal rapid evolution by wild honey bees exposed to a novel parasite. *Nat Commun, 6*, 7991. doi: 10.1038/ncomms8991

Navajas, M., Migeon, A., Alaux, C., Martin-Magniette, M. L., Robinson, G. E., Evans, J. D., Cros-Arteil, S., Crauser, D., & Le Conte, Y. (2008). Differential gene expression of the honey bee *Apis mellifera* associated with *Varroa destructor* infection. *BMC Genomics, 9*(1 1471-2164), 301. doi: 10.1186/1471-2164-9-301

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., & Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome Res, 15*(11), 1566-1575. doi: 10.1101/gr.4252305

Peñuelas, J., & Filella, I. (2001). Responses to a Warming World. *Science, 294*(5543), 793-795. doi: 10.1126/science.1066860

Prunier, J., Laroche, J., Beaulieu, J., & Bousquet, J. (2011). Scanning the genome for gene SNPs related to climate adaptation and estimating selection at the molecular level in boreal black spruce. *Mol Ecol, 20*(8), 1702-1716. doi: 10.1111/j.1365-294X.2011.05045.x

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet, 81*(3), 559-575. doi: 10.1086/519795

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Mol Ecol, 24*(17), 4348-4370. doi: 10.1111/mec.13322

Ruttner, F. (1988). *Biogeography and taxonomy of honeybees*. Berlin: Springer Verlag.

Sandoz, J. C. (2011). Behavioral and neurophysiological study of olfactory perception and learning in honeybees. *Front Syst Neurosci, 5*, 98. doi: 10.3389/fnsys.2011.00098

Schlamp, F., van der Made, J., Stambler, R., Chesebrough, L., Boyko, A. R., & Messer, P. W. (2016). Evaluating the performance of selection scans to detect selective sweeps in domestic dogs. *Mol Ecol, 25*(1), 342-356. doi: 10.1111/mec.13485

Schowalter, T. D. (2016). *Insect ecology: an ecosystem approach*: *Academic Press*.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Res, 33*(Web Server issue), W382-388. doi: 10.1093/nar/gki387

Sheppard, W. S., & Meixner, M. D. (2003). *Apis mellifera pomonella*, a new honey bee subspecies from Central Asia. *Apidologie, 34*(4), 367-375.

Simpson, J. (1958). The problem of swarming in beekeeping practice. *Bee World, 39*(8), 193-202. doi: 10.1080/0005772x.1958.11095063

Sojo, V., Dessimoz, C., Pomiankowski, A., & Lane, N. (2016). Membrane proteins are dramatically less conserved than water-soluble proteins across the Tree of Life. *Mol Biol Evol, 33*(11), 2874-2884. doi: 10.1093/molbev/msw164

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A, 100*(16), 9440-9445. doi: 10.1073/pnas.1530509100

Strange, J., P., Garnery, L., & Sheppard, W., S. (2007). Persistence of the Landes ecotype of *Apis mellifera mellifera* in southwest France: confirmation of a locally adaptive annual brood cycle trait. *Apidologie, 38*(3), 259-267.

Strange, J. P., Garnery, L., & Sheppard, W. S. (2008). Morphological and molecular characterization of the Landes honey bee (*Apis mellifera L.*) ecotype for genetic conservation. *Journal of Insect Conservation, 12*(5), 527-537. doi: 10.1007/s10841-007-9093-6

Stucki, S., Orozco-terWengel, P., Bruford, M. W., Colli, L., Masembe, C., Negrini, R., Taberlet, P., Joost, S., & Consortium, N. (2014). High performance computation of landscape genomic models integrating local indices of spatial association. *arXiv:1405.7658v1*.

Sun, L., Liu, S., Wang, R., Jiang, Y., Zhang, Y., Zhang, J., Bao, L., Kaltenboeck, L., Dunham, R., Waldbieser, G., & Liu, Z. (2014). Identification and analysis of genome-wide SNPs provide insight into signatures of selection and domestication in channel catfish (*Ictalurus punctatus*). *PLoS One, 9*(10), e109666. doi: 10.1371/journal.pone.0109666

Szpiech, Z. A., & Hernandez, R. D. (2014). selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol Biol Evol, 31*(10), 2824-2827. doi: 10.1093/molbev/msu211

Tauber, E., Zordan, M., Sandrelli, F., Pegoraro, M., Osterwalder, N., Breda, C., Daga, A., Selmin, A., Monger, K., Benna, C., Rosato, E., Kyriacou, C. P., & Costa, R. (2007). Natural selection favors a newly derived *timeless* allele in *Drosophila melanogaster*. *Science, 316*(5833), 1895-1898. doi: 10.1126/science.1138412

Thioulouse, J., Chessel, D., Doledec, S., & Olivier, J. M. (1997). ADE-4: a multivariate analysis and graphical display software. *Statistics and Computing, 7*, 75-83.

Triska, P., Soares, P., Patin, E., Fernandes, V., Cerny, V., & Pereira, L. (2015). Extensive admixture and selective pressure across the Sahel Belt. *Genome Biol Evol, 7*(12), 3484-3495. doi: 10.1093/gbe/evv236

Vasemagi, A., & Primmer, C. R. (2005). Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. *Mol Ecol, 14*(12), 3623-3642. doi: 10.1111/j.1365-294X.2005.02690.x

Vernot, B., Stergachis, A. B., Maurano, M. T., Vierstra, J., Neph, S., Thurman, R. E., Stamatoyannopoulos, J. A., & Akey, J. M. (2012). Personal and population genomics of human regulatory variation. *Genome Res, 22*(9), 1689-1697. doi: 10.1101/gr.134890.111

Voight, B. F., Kudaravalli, S., Wen, X., & Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biol, 4*(3), e72. doi: 10.1371/journal.pbio.0040072

Wallberg, A., Han, F., Wellhagen, G., Dahle, B., Kawata, M., Haddad, N., Simoes, Z. L., Allsopp, M. H., Kandemir, I., De la Rua, P., Pirk, C. W., & Webster, M. T. (2014). A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee *Apis mellifera*. *Nat Genet, 46*(10), 1081-1088. doi: 10.1038/ng.3077

Wallberg, A., Pirk, C. W., Allsopp, M. H., & Webster, M. T. (2016). Identification of multiple loci associated with social parasitism in honeybees. *PLoS Genet, 12*(6), e1006097. doi: 10.1371/journal.pgen.1006097

Wang, X., Liu, J., Zhou, G., Guo, J., Yan, H., Niu, Y., Li, Y., Yuan, C., Geng, R., Lan, X., An, X., Tian, X., Zhou, H., Song, J., Jiang, Y., & Chen, Y. (2016). Whole-genome sequencing of eight goat populations for the detection of selection signatures underlying production and adaptive traits. *Scientific Reports, 6*, 38932. doi: 10.1038/srep38932

Waples, R. S., & Gaggiotti, O. (2006). What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Mol Ecol, 15*(6), 1419-1439. doi: 10.1111/j.1365-294X.2006.02890.x

Watanabe, A., Choe, S., Chaptal, V., Rosenberg, J. M., Wright, E. M., Grabe, M., & Abramson, J. (2010). The mechanism of sodium and substrate release from the binding pocket of vSGLT. *Nature, 468*(7326), 988-991. doi: 10.1038/nature09580

Weiss, S., & Ferrand, N. X. (2007). *Phylogeography of southern European refugia*: Springer.

Wiegand, A., Stucki, L., Hoffmann, R., Attin, T., & Stawarczyk, B. (2015). Repairability of CAD/CAM high-density PMMA- and composite-based polymers. *Clin Oral Investig, 19*(8), 2007-2013. doi: 10.1007/s00784-015-1411-x

Wray, G. A. (2007). The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet, 8*(3), 206-216. doi: 10.1038/nrg2063

Xia, J. H., Bai, Z., Meng, Z., Zhang, Y., Wang, L., Liu, F., Jing, W., Wan, Z. Y., Li, J., Lin, H., & Yue, G. H. (2015). Signatures of selection in tilapia revealed by whole genome resequencing. *Scientific Reports, 5*, 14168. doi: 10.1038/srep14168

Xu, K., DiAngelo, Justin R., Hughes, Michael E., Hogenesch, John B., & Sehgal, A. (2011). The circadian clock interacts with metabolic physiology to influence reproductive fitness. *Cell Metabolism, 13*(6), 639-654. doi: 10.1016/j.cmet.2011.05.001

Xu, X., Gopalacharyulu, P., Seppänen-Laakso, T., Ruskeepää, A.-L., Aye, C. C., Carson, B. P., Mora, S., Orešič, M., & Teleman, A. A. (2012). Insulin signaling regulates fatty acid catabolism at the level of CoA activation. *PLoS Genet, 8*(1), e1002478. doi: 10.1371/journal.pgen.1002478

Yerushalmi, S., & Green, R. M. (2009). Evidence for the adaptive significance of circadian rhythms. *Ecology Letters, 12*(9), 970-981. doi: 10.1111/j.1461-0248.2009.01343.x

Yon Rhee, S., Wood, V., Dolinski, K., & Draghici, S. (2008). Use and misuse of the gene ontology annotations. *Nat Rev Genet, 9*(7), 509-515. doi: 10.1038/nrg2363

Yuan, Q., Lin, F., Zheng, X., & Sehgal, A. (2005). Serotonin modulates circadian entrainment in Drosophila. *Neuron, 47*(1), 115-127. doi: 10.1016/j.neuron.2005.05.027

Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res, 33*(7), 2302-2309. doi: 10.1093/nar/gki524

Zueva, K. J., Lumme, J., Veselov, A. E., Kent, M. P., Lien, S., & Primmer, C. R. (2014). Footprints of directional selection in wild Atlantic salmon populations: evidence for parasite-driven evolution? *PLoS One, 9*(3), e91672. doi: 10.1371/journal.pone.0091672

Chapter VI.

Developing reduced SNP assays from whole-genome sequence data to estimate introgression in an organism with complex genetic patterns, the Iberian honey bee (*Apis mellifera iberiensis*)

The paper was submitted to the Journal *Evolutionary Applications*:

Dora Henriques, Melanie Parejo , Alain Vignal, David Wragg, Andreas Wallberg, Matthiew Tomas Webster and Maria Alice Pinto

Abstract

The most important managed pollinator, the honey bee (*Apis mellifera* L.), has been subject to a growing number of threats. In Western Europe one such threat is large-scale introductions of commercial strains (C-lineage ancestry), which is leading to introgressive hybridization and even the local extinction of native honey bee populations (M-lineage ancestry). Here, we developed reduced assays of highly informative SNPs from 176 whole genomes to estimate C-lineage introgression in the most diverse and evolutionarily complex subspecies in Europe, the Iberian honey bee (*Apis mellifera iberiensis*). We started by evaluating the effects of sample size and sampling a geographically restricted area on the number of highly informative SNPs. We demonstrated that a bias in the number of fixed SNPs (F_{ST}=1) is introduced when the sample size is small (N≤10) and when sampling only captures a small fraction of a population's genetic diversity. These results underscore the importance of having a representative sample when developing reliable reduced SNP assays for organisms with complex genetic patterns. We used a training dataset to design four independent SNP assays selected from pairwise FST between the Iberian and C-lineage honey bees. The designed assays, which were validated in holdout and simulated hybrid datasets, proved to be highly accurate and can be readily used for monitoring populations not only in the native range of *A. m. iberiensis* in Iberia but also in the introduced range in the Balearic islands, Macaronesia, and South America, in a time- and cost-effective manner. While our approach used the Iberian honey bee as model system, it has a high value in a wide range of scenarios for the monitoring and conservation of potentially hybridized domestic and wildlife populations.

Keywords: *Apis mellifera iberiensis*, F_{ST}, informative SNPs, reduced SNP assays

Introduction

Biodiversity, including the genetic diversity within and between populations, is a unique heritage whose conservation is imperative for the benefit of future generations (Frankham *et al.*, 2002). This is particularly important for organisms like the honey bee (*Apis mellifera* L.), which, through the pollination service it provides, plays a critical role in ecosystem functioning and in food production for humanity. The honey bee is under pressure worldwide due to multiple factors, ranging from emergent parasites and pathogens, and the overuse of agrochemicals, to the less publicized introgressive hybridization mediated by human management (reviewed by Potts *et al.*, 2010; vanEngelsdorp & Meixner, 2010). In a global world, where the circulation of commercial queens and package honey bees occurs at a rapid pace, and at large scale, reliable tools for monitoring genetic diversity are becoming indispensable.

The honey bee exhibits high diversity, with 31 currently recognized subspecies (Chen *et al.*, 2016; Engel, 1999; Meixner *et al.*, 2011; Sheppard & Meixner, 2003) belonging to four main evolutionary lineages (Western and Northern Europe, M; Southeastern Europe, C; Africa, A; Middle East and Central Asia, O). Of the 31 subspecies, the Iberian honey bee *A. m. iberiensis* (M-lineage) has received the most attention with numerous genetic surveys (Chávez-Galarza *et al.*, 2015, and references therein). These have consistently shown the existence of a highly diverse and structured subspecies defined by two major clusters forming a sharp cline that bisects Iberia along a northeastern–southwestern axis (Arias *et al.*, 2006; Chávez-Galarza *et al.*, 2017; Smith *et al.*, 1991). Such complexity has been shaped by recurrent cycles of interacting selective and demographic processes, typical of long-term glacial refugia organisms (Chávez-Galarza *et al.*, 2017; Chávez-Galarza *et al.*, 2013; Chávez-Galarza *et al.*, 2015). However, this genetic legacy might be at risk if Iberian beekeepers adopt a strategy of importing commercial strains belonging to the highly divergent lineage C, as is occurring at large scale throughout Western and Northern Europe North of the Pyrenees. Since the early 20th century, beekeeping activity in this part of Europe has been characterized by colony importations and queen breeding with mostly C-lineage honey bees (De la Rúa *et al.*, 2009), which are known for their docile nature and high productivity (Ruttner, 1988). This human-mediated gene flow has threatened *A. m. mellifera,* the other M-lineage subspecies besides *A. m. iberiensis* in Europe. Indeed, the genetic integrity of *A. m. mellifera* has been compromised by introgressive hybridization and in some areas it has even been replaced by subspecies of C-lineage ancestry (Jensen *et al.*, 2005; Pinto *et al.*, 2014; Soland-

Reckeweg *et al.*, 2009). Yet, maintaining locally adapted subspecies is crucial for the long-term sustainability of *A. mellifera* (De la Rúa *et al.*, 2013; vanEngelsdorp & Meixner, 2010). Reciprocal translocation experiments have recently shown that local honey bees have longer survivorship (Büchler *et al.*, 2014) and lower pathogen loads (Francis *et al.*, 2014) than introduced ones, reinforcing the importance of preserving the genetic diversity of locally adapted subspecies. Furthermore, it has been advocated that apiculture and commercial breeding could compromise honey bee health by interfering with natural selection (Meixner *et al.*, 2010; Neumann & Blacquière, 2017).

The idea that long-term sustainability of honey bee populations can only be achieved by preserving natural genetic diversity and co-evolved gene complexes has led to the establishment of conservation programs and protected areas throughout Europe (De la Rúa *et al.*, 2009). To foster and monitor such conservation efforts, reliable, cost- and time-effective tools are needed to accurately assess admixture levels between introduced and native honey bees. For the endangered *A. m. mellifera*, reduced assays of highly informative SNPs have already been developed to estimate C-lineage introgression (Muñoz *et al.*, 2015; Parejo *et al.*, 2016). However, equivalent tools for application in conservation and breeding efforts are still required for its sister subspecies, *A. m. iberiensis*.

Following the last glacial maximum, honey bees dispersed from the Iberian refugium to colonize a broad territory, extending from the Pyrenees to the Urals (Franck *et al.*, 1998; Ruttner, 1988). This important Iberian reservoir of genetic diversity has not yet been seriously threatened by C-lineage introgression (Chávez-Galarza *et al.*, 2017; Chávez-Galarza *et al.*, 2015; Miguel *et al.*, 2007), although this scenario might change as many young beekeepers are attracted by the advertised benefits of commercial strains – being more prolific and docile. In many islands of the Baleares and Macaronesia, for example, where the Iberian honey bee was presumably introduced in historical times, the contemporaneous large-scale importation of commercial C-lineage queens has resulted in high levels of introgression into the local populations (De La Rúa *et al.*, 2001; De la Rúa *et al.*, 2003; Miguel *et al.*, 2015; Muñoz *et al.*, 2014). The conservation of *A. m. iberiensis* diversity is therefore a priority, especially in the light of climate change as this subspecies is well adapted to a broad range of environments, including hot and dry summer months with limited nectar flows. These adaptations could be a basis for selection of new development cycles suited to new environmental conditions (Le Conte & Navajas, 2008).

A diverse array of molecular tools has been employed to monitor C-lineage introgression including PCR-RFLP of the intergenic $tRNA^{leu}$-cox2 mtDNA region (Bertrand *et al.*, 2015), microsatellites (Jensen *et al.*, 2005; Soland-Reckeweg *et al.*, 2009), and more recently SNPs (Parejo *et al.*, 2016; Pinto *et al.*, 2014). Among these, SNPs are becoming the tool of choice for many applications because they are easily transferred between laboratories, have low genotyping error, provide high quality data, are suitable for automation in high throughput technologies (Vignal *et al.*, 2002), and are more powerful for estimating introgression in honey bees (Muñoz *et al.*, 2017).

High-throughput sequencing of whole-genomes generates millions of SNPs. Yet, this volume of data is inappropriate for routine conservation purposes, such as breeding and population monitoring. Therefore, the mining of highly informative SNPs from such high genomic resolution datasets is a common approach for developing reduced SNP assays capable of reliable ancestry estimation (Amirisetty *et al.*, 2012; Judge *et al.*, 2017). While different metrics and approaches (e.g. Delta, I_n, PCA, outlier tests) can be used for ranking SNPs by information content, the Fixation index (F_{ST}) has been the metric of choice perhaps due to its power (Ding *et al.*, 2011; Karlsson *et al.*, 2011; Wilkinson *et al.*, 2011), especially when comparing only two highly divergent populations (Hulsegge *et al.*, 2013). Furthermore, some metrics are correlated regarding information content, in particular those based on allele frequencies (Ding *et al.*, 2011; Wilkinson *et al.*, 2011).

In this study, we developed cost-effective reduced SNP assays from 176 whole-genome sequences. When developing such tools, to assure that they are accurate and reliable, the diversity and population complexity needs to be considered. Therefore, taking advantage of the large and comprehensive whole-genome dataset for *A. m. iberiensis* (N=117), we first tested the effect of sample size and sampling a geographically restricted area on detecting fixed SNPs. Next, we designed the reduced SNP assays using a training dataset to identify highly informative SNPs (F_{ST}=1), which were then validated in holdout and simulated datasets. The constructed SNP assays were revealed to be very powerful for accurately estimating C-lineage introgression and can thus be applied to support conservation efforts in the Iberian honey bee.

Material and Methods

Samples

The whole-genome sequences used in this study were obtained from 176 pure haploid males, representing 117 *A. m. iberiensis*, 28 *A. m. carnica* and 31 *A. m. ligustica* (DH and MAP, unpublished data; Parejo *et al.*, 2016; Wragg et al., in preparation) sampled across a wide geographical range (Figure VI-1). All samples were sequenced on an Illumina HiSeq 2500 with an aimed sequencing depth of 10X per individual. Mapping and variant calling was performed following best practices (see Supporting Information for details).

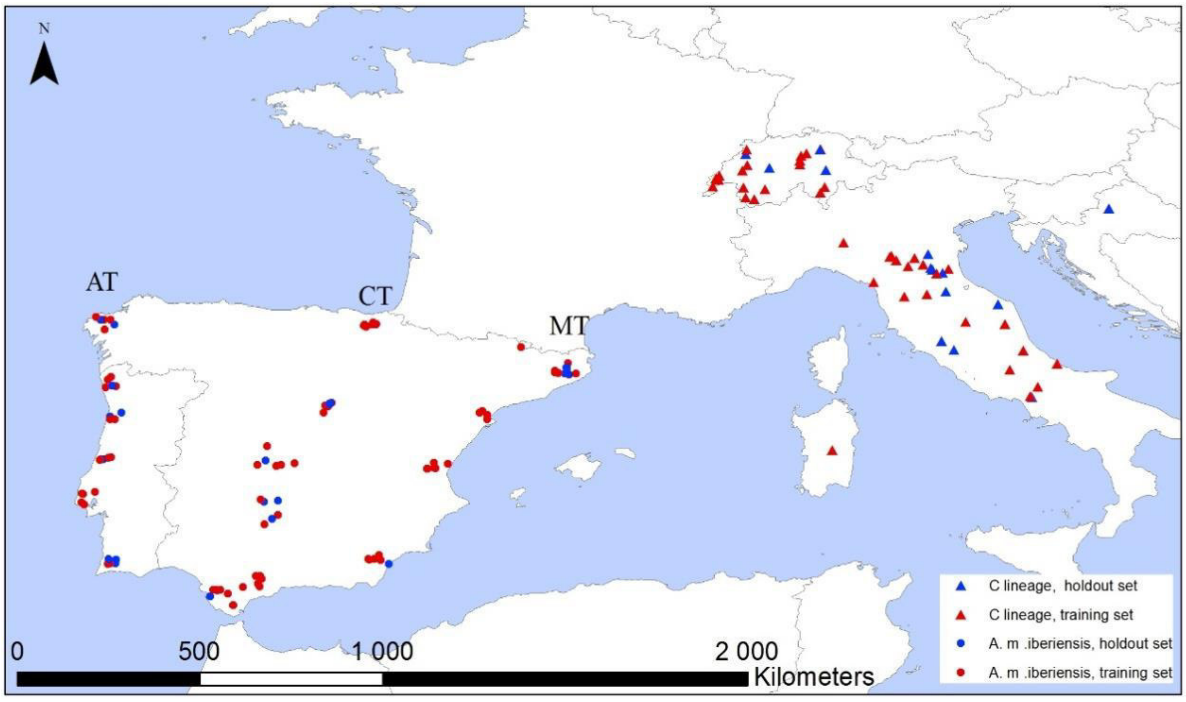


Figure VI-1 - Geographic locations of the 176 whole-genome sequenced individuals. The Iberian honey bees are distributed across the three transects: Atlantic (AT; N=31), Central (CT; N=61), and Mediterranean (MT, N=25). Each dot represents a single colony and apiary

To assess subspecies ancestry and purity of all individuals included in the initial whole-genome dataset (see Supporting Information for details), we inferred model-based admixture proportions (Q-values) for K=1 to 5 clusters with 10,000 iterations using the software ADMIXTURE v1.3.0 (Alexander *et al.*, 2009). We employed Q-value thresholds of >0.95 and <0.05 for defining subspecies ancestry and purity of C-lineage and M-lineage subspecies, respectively. Convergence between independent runs was monitored by comparing the resulting log-likelihood scores (LLS)

using the default termination criterion set to stop when LLS increases by less than 0.0001 between runs. The optimal number of K clusters was determined using cross-validation (CV) error as implemented in ADMIXTURE. Q-values were visualised in R (R Core Team, 2016). To have an overall estimate on population divergence, we calculated in PLINK 1.9 (Chang *et al.*, 2015) the average genome-wide pairwise F_{ST} (Weir & Cockerham, 1984) between *A. m. iberiensis, A. m. carnica*, and *A. m. ligustica* and between *A. m. iberiensis* and combined *A. m. carnica* with *A. m. ligustica* (C-lineage).

Effect of sampling bias on the number of fixed SNPs

Starting with a large sample size, which covers a species' entire geographical range and therefore encompasses its variation, is an important first step for developing SNP assays with high statistical power (Ding *et al.*, 2011; Mariette *et al.*, 2002). Using the large (N=117) and geographically comprehensive sample of *A. m. iberiensis* (Figure VI-1), we assessed the effects of sample size and of sampling a geographically restricted area on the number of fixed SNPs (F_{ST}=1).

To test the effect of sample size, we constructed 30 subsets with different sample sizes (N=5, 10, 25, 50, 75 and 100, five replicates each) by randomly choosing individuals from the complete dataset (N=117) of *A. m. iberiensis* (Figure VI-2). Next, we calculated the number of fixed SNPs between each of the 30 *A. m. iberiensis* subsets and the C-lineage dataset (N=59) using PLINK. The number of fixed SNPs identified for each replicate was subtracted from the number of fixed SNPs calculated with the complete *A. m. iberiensis* dataset. This approach provided an estimate of the number of SNPs erroneously identified as fixed between the two groups, due to limited sampling effort (false positive fixed SNPs).

To test the effect of sampling a geographically restricted area, we constructed four different subsets by randomly choosing 25 individuals (N=25) from the following areas: Portugal (PT; this sample may arise in practice when sampling is country-limited), Central transect (CT; sampling representing the largest latitudinal distance in Iberia), Mediterranean transect (MT; sampling along the Mediterranean coast mimics the pioneer mtDNA survey carried out by Smith *et al.* (1991), and across the Iberian Peninsula (IP) to intentionally capturing the entire variation in *A. m. iberiensis*. The number of fixed SNPs between the C-lineage dataset (N=59) and each of the four subsets was subtracted from the number of fixed SNPs calculated with the complete *A. m. iberiensis* dataset.

The number of false positive fixed SNPs was then compared among the four subsets (Figure VI-2).

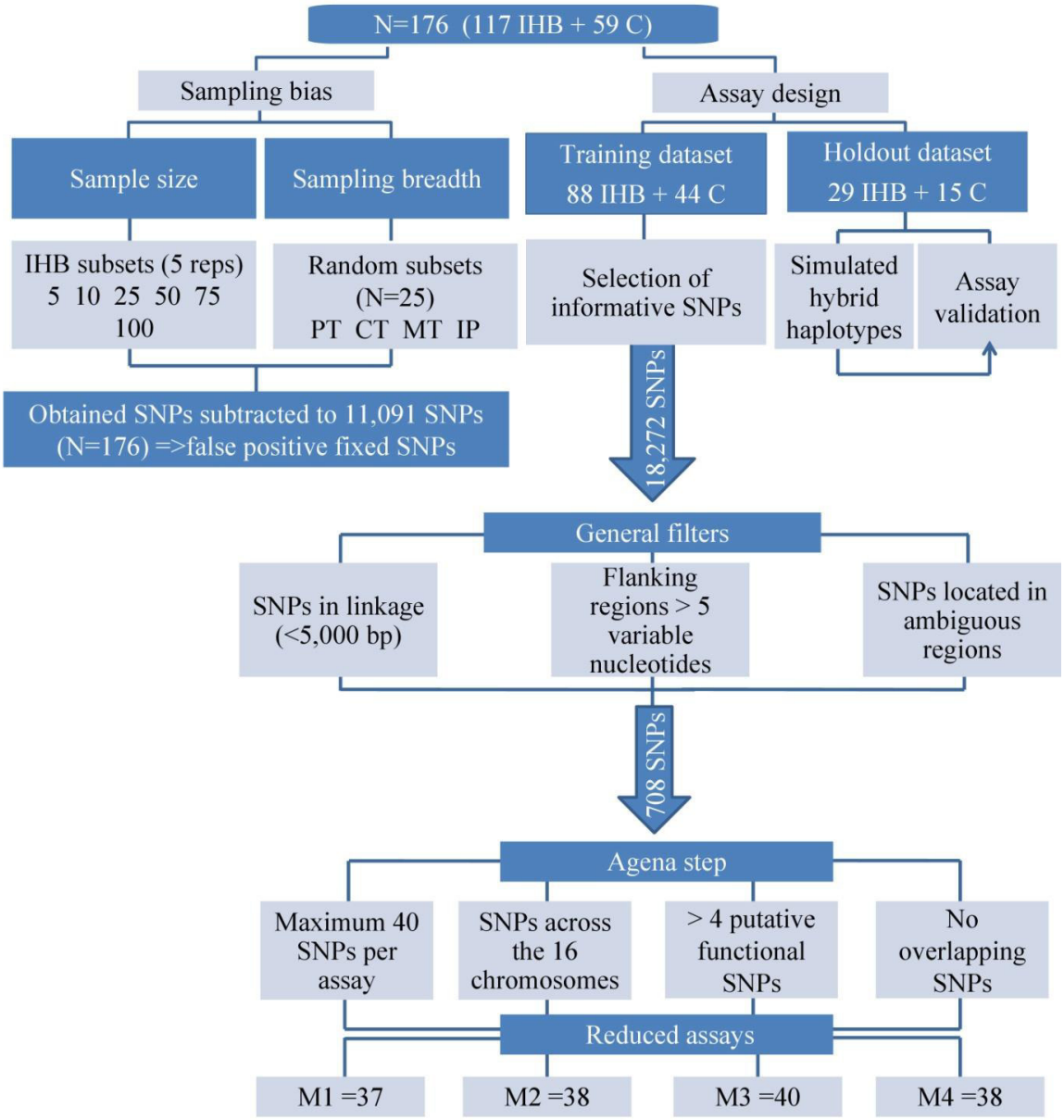


Figure VI-2 - Diagram depicting the different phases of development of the four reduced SNP assays (M1, M2, M3, M4) using as a baseline whole-genome sequence data from 117 *A. m. iberiensis* (IHB) and 59 C-lineage

Assay design

After assessing the effects of sampling bias on the number fixed SNPs, we proceeded with designing the reduced SNP assays for estimating C-lineage introgression into *A. m. iberiensis* (Figure VI-2). We followed Anderson's simple training and holdout method to minimise the bias which is introduced when selection and assessment of informative SNPs are based on the same individuals (Anderson, 2010). Accordingly, we set aside a holdout dataset, consisting of 29 *A. m.*

iberiensis and 15 C-lineage individuals chosen at random (25% of the total sample size), for subsequent assay validation (Table VI-1). The remaining 88 *A. m. iberiensis* and 44 C-lineage individuals (23 *A. m. carnica*, 21 *A. m. ligustica*) were used as the training dataset for selecting informative SNPs.

Table VI-1 - Sample sizes of training and holdout datasets for each population

Population	Training set	Holdout set	Total
A. m. iberiensis	88	29	117
C-lineage *(A. m. carnica & A. m. ligustica)*	44 (23 +21)	15 (5 + 10)	59 (28 + 31)
Total	132	44	176

The most informative SNPs were identified from F_{ST} values (fixed SNPs, F_{ST}=1), calculated in PLINK between *A. m. iberiensis* and C-lineage individuals using the training dataset. To uncover the putative functional role of the highly differentiated SNPs, we used SNPeff 4.3 (Cingolani *et al.*, 2012) and the NCBI honey bee annotation version 102 (Pruitt *et al.*, 2013). Subsequently, we performed a Gene Ontology (GO) analysis in the DAVID v.8.0 database (Huang *et al.*, 2009) considering the GO terms of the Biological Process (BP), Molecular Function (MF), Cellular Component (CC) (Gene Ontology Consortium, 2015), and the KEGG pathway (Kanehisa *et al.*, 2016).

To downsize the number of fixed SNPs, the first filter eliminated SNPs <5,000 bp apart, which carry redundant information (Figure VI-2). This distance threshold correlates with the high linkage disequilibrium (LD) decay in honey bees (Wallberg *et al.*, 2015) and has been used by others (Chapman *et al.*, 2015; Harpur *et al.*, 2014). In this filtering step, SNPs located in 3'UTR, 5'UTR, missense, splice donor, and splice regions were preferentially retained to assure that the reduced assays included SNPs of putative functional relevance and thereby represent real phenotypic differences between lineages.

The subsequent filtering step was linked to the Agena Bioscience MassARRAY® MALDI-TOF genotyping system (Figure VI-2). To increase the probability of amplification success, we removed the SNPs which had >5 variable nucleotides on either side of the 250 bp flanking sequences, which will be used for primer design (Table Sup VI-1). Additionally, SNPs located in ambiguous regions of the reference genome were excluded using the following criteria: (i) >5 sequential unknown nucleotides (N) in the flanking regions, (ii) flanking regions matching multiple

contigs on the reference genome, and (iii) flanking regions consisted of short repeats. The remaining SNPs were used to design four multiplexes (M1, M2, M3, M4) with the software Assay Design 4.0 (www.agenabio.com), which selects the best combination of SNPs for amplification by preventing hairpin and dimer formation. Three criteria were followed to construct each multiplex (hereafter termed reduced SNP assay) aiming at a maximum of 40 SNPs per multiplex, as allowed by the MassARRAY® technology: (i) every chromosome represented, (ii) at least four putative functional SNPs, and (iii) no overlapping SNPs between multiplexes. For comparison purposes, we also constructed four assays of randomly chosen SNPs (hereafter termed random SNP assays) from the whole-genome dataset with the same size of the four multiplexes.

Assay Validation

For validating the reduced SNP assays, we simulated hybrid haplotypes using the software admix-simu (https://github.com/williamslab/admix-simu) and a window-based 100 kbp resolution recombination map from Wallberg *et al.* (2015). To avoid related haplotypes in the simulated F1 and backcross haplotypes, we used the parental individuals only once in the simulation of recombination. The 29 *A. m. iberiensis* and the 15 C-lineage individuals of the holdout dataset were randomly chosen to simulate the hybrid haplotypes as follows: F1s were simulated using 15 *A. m. iberiensis* and 15 C-lineage individuals as parents; backcrosses were simulated using 14 F1 and the remaining 14 *A. m. iberiensis* individuals as parents.

The reduced and random SNP assays were validated in the holdout (N=44) and simulated datasets (N=29) by estimating the Q-values with ADMIXTURE, using the unsupervised option and the default settings, for K=2 and 200 bootstrap replicates. We examined the performance of each reduced and random SNP assay (individually or by combining the best performing assays) against the whole-genome dataset, which provides the true Q-value, by calculating (i) deviation, (ii) precision and (iii) accuracy. Precision was assessed by the Pearson correlation coefficient (r) and the standard deviation of the differences. Accuracy was assessed through the percentage of absolute error.

Results

SNP calling and population structure

A total of 2,366,382 SNPs were detected in the whole-genome sequences of 176 individuals (117 *A. m. iberiensis*, 31 *A. m. ligustica*, 28 *A. m. carnica*), with a genotyping rate of 0.986. Information on sample origin, coverage and variant calling statistics is provided in Table Sup VI-2.. Using the whole-genome sequences, the global pairwise F_{ST} values were estimated for the M-lineage *A. m. iberiensis* and the C-lineage *A. m. carnica* and *A. m. ligustica* (Table VI-2). As expected, F_{ST} between the subspecies belonging to the highly divergent M and C lineages was high (F_{ST}≥0.53) whereas between the closely related *A. m. carnica* and *A. m. ligustica* was low (F_{ST}=0.06). The two lineages are clearly separated at the optimal K=2 (Fig. Sup VI-1), with the 117 *A. m. iberiensis* individuals forming one cluster, and the 28 *A. m. carnica* together with the 31 *A. m. ligustica* individuals forming another cluster (Figure Sup VI-2).

Table VI-2 - Population differentiation estimated from average genome-wide F_{ST}

Population	***A. m. carnica***	***A. m. ligustica***	**C-lineage (*A. m. carnica* & *A. m. ligustica*)**
A. m. iberiensis	0.540	0.549	0.532
A. m. ligustica	0.061		

Effect of sampling bias on the number of fixed SNPs

The effect of sample size and sampling a geographically restricted area on the number of fixed SNPs (F_{ST}=1) was examined to understand to what extent false positive fixed SNPs would bias reduced SNP assays for estimating introgression. A total of 11,091 fixed SNPs were detected between the complete *A. m. iberiensis* dataset (N=117) and the C-lineage dataset (N=59). As expected, the number of fixed SNPs and the number of false positives increases as the *A. m. iberiensis* sample size decreases, and this trend is more pronounced when N<25 (Table VI-3). For N=5, a large proportion of false positives (33.9%) displayed a FST ≤0.95 with a minimum of 0.084, which might impact the power of reduced SNP assays. However, the impact is negligible for N≥25 as the proportion of false positives is ≤3.4% and the minimum FST value (0.695) is still relatively high (Table VI-3).

Table VI-3 - Fixed SNPs and 95% confidence interval (CI) estimated from random subsets of variable sample size (5 replicates each) of *A. m. iberiensis* and statistics for F_{ST} values estimated from the false positive fixed SNPs

Sample size subset	Mean number of fixed SNPs (± 95 % CI)	Mean number of false positive fixed SNPs*	Mean % of false positive fixed SNPs with an $F_{ST} \leq 0.95$**	Mean minimum F_{ST}
5	25.428 (± 1.184)	14.337	33.9	0.084
10	18.878 (± 354)	7.787	14	0.334
25	15.700 (± 127)	4.609	3.4	0.695
50	13.784 (± 282)	2.693	0.3	0.880
75	12.480 (± 306)	1.389	0.1	0.942
100	11.736 (± 165)	645	0	0.970

*Calculated by subtracting the number of fixed SNPs estimated for each sample size subset from 11,091 fixed SNPs estimated for the complete dataset of *A. m. iberiensis* (N=117), which displays a minimum F_{ST}=1.
**Calculated by retrieving the F_{ST} values obtained from the complete *A. m. iberiensis* dataset for the false positives and calculating the percentage with a $F_{ST} \leq 0.95$.

Sampling a geographically restricted area also influences the number of fixed SNPs, although the extent of bias depends on sample origin (Table VI-4). Interestingly, the highest number of false positives is identified when sampling is restricted to Portugal (PT). In contrast, sampling along the north-south transect in the centre of Iberia (CT) provides the best estimate of fixed SNPs. Considering the percentage of false positives with a $F_{ST} \leq 0.95$, the best result was obtained for the IP subset with only 10.4% and with a minimum value of F_{ST}=0.763. This contrasted with the PT subset for which there were twice as many (20.2%) false positives with a $F_{ST} \leq 0.95$ and a considerably lower minimum value of 0.275 (Table VI-4).

Table VI-4 - Fixed SNPs estimated from geographical subsets of *A. m. iberiensis* and statistics for F_{ST} values estimated from the false positive fixed SNPs

Geographical subset*	Number of fixed SNPs	Number of false positive fixed SNPs**	% of false positive fixed SNPs with an $F_{ST} \leq 0.95$***	Minimum F_{ST}
PT	17,738	6,647	20.2	0.275
CT	15,009	3,918	13.7	0.700
MT	15,384	4,293	11.8	0.676
IP	15,371	4,280	10.4	0.763

*PT: Portugal, CT: Central transect, MT: Mediterranean transect and IP: Iberian Peninsula.
** Calculated by subtracting the number of fixed SNPs estimated for each geographical subset from 11,091 fixed SNPs estimated for the complete dataset of *A. m. iberiensis* (N=117), which displays a minimum F_{ST}=1.
***Calculated by retrieving the F_{ST} values obtained from the complete *A. m. iberiensis* dataset for the false positives and calculating the percentage with a $F_{ST} \leq 0.95$.

Selection and genomic information of highly informative SNPs

Having assessed the potential effects of sampling bias we were able to follow Anderson's simple training and holdout method without incorporating a significant bias when selecting highly informative SNPs (Figure VI-2). Accordingly, highly informative SNPs for estimating C-lineage introgression into *A. m. iberiensis* were selected using the training dataset (88 *A. m. iberiensis* and 44 C-lineage individuals). A total of 18,272 SNPs were fixed (F_{ST}=1) (Table Sup VI-3, Figure Sup VI-3), an increase of 7,181 fixed SNPs compared to that calculated from the complete dataset (117 *A. m. iberiensis* dataset and 59 C-lineage individuals). While these SNPs were not fixed in the complete dataset, they were still highly differentiated (F_{ST}≥0.95 for 98.9% of the SNPs; minimum F_{ST}=0.925) and thereby highly informative.

The 18,272 SNPs were distributed across the 16 honey bee chromosomes (Figure Sup VI-3), and located in 247 intergenic regions and 1,347 genic regions (±5 kb around coding sequences; Table Sup VI-3). Chromosome 11 contained the highest proportion of fixed SNPs (3.1%, 4,729 SNPs) whereas chromosome 7 had the least (0.3%, 400 SNPs; Table Sup VI-4). The physical distance between the fixed SNPs ranged from 1 bp to 2,587,074 bp with a mean of 11,261 bp. Most fixed SNPs are located in introns (7,666) and intergenic regions (4,257), however a number are located in regions of putative functional relevance, including 47 SNPs (distributed along 37 genes) that are non-synonymous or missense variants (Table Sup VI-5). Of the 1,347 genic regions containing SNPs, 12 harbour more than 100 SNPs (Table Sup VI-6). Gene ontology (GO) analysis revealed 13 significantly enriched functional terms (modified Fisher exact P-value<0.05; Table Sup VI-7). The Biological Processes term "regulation of transcription, DNA-templated" shared 12 genes with the Molecular Function term, "transcription factor activity, sequence-specific DNA binding". Two other Molecular Function terms are associated with more than 26 genes related to DNA binding ("sequence-specific DNA binding", "DNA binding"). The KEGG pathways were represented by four terms "Aminoacyl-tRNA biosynthesis", "Wnt signalling pathway", "mRNA surveillance pathway" and "Insulin resistance".

Assay design

Several filters were applied to the initial 18,272 fixed SNPs identified in the training dataset resulting in a final dataset of 708 SNPs, which were used to design four multiplexes (or reduced assays) with the Assay Design tool of Agena (Figure VI-2). The resulting assays contained 37 (M1),

38 (M2), 40 (M3), and 38 (M4) SNPs. Each assay combines highly informative SNPs covering 15 (M1 lacks SNPs in chromosome 16, M2 in chromosome 14) or 16 (M3, M4) chromosomes (Figure VI-3, Table Sup VI-4).

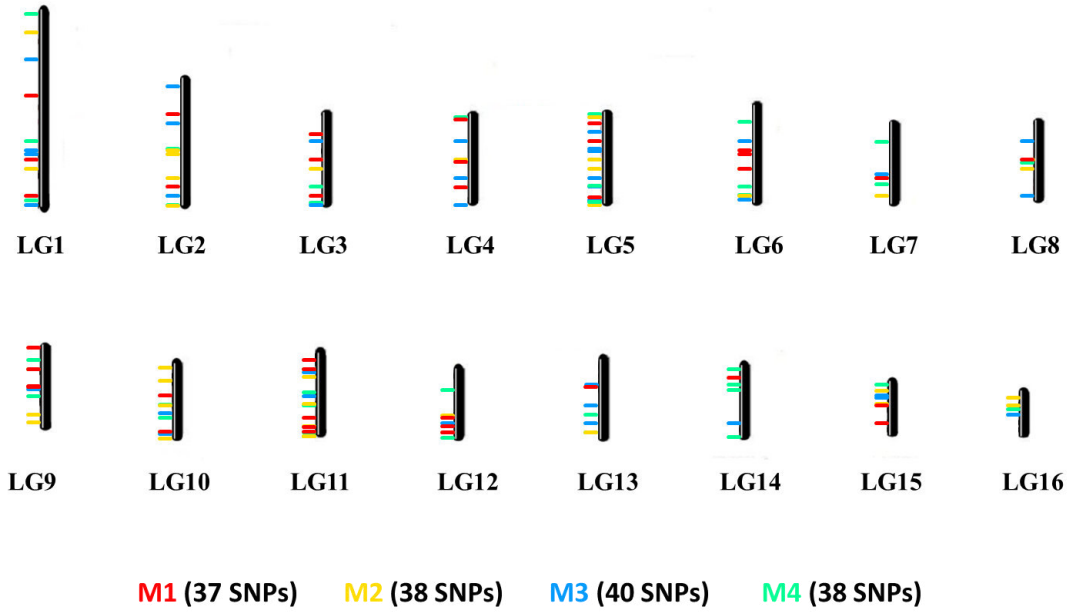


Figure VI-3 - Chromosome map showing the SNP positions of the four reduced assays (M1-M4)

Assay validation

The reduced (M1, M2, M3, M4) and random SNP assays (R1, R2, R3, R4) were validated in the holdout (29 *A. m. iberiensis*) and simulated (29 hybrid haplotypes) datasets (Figure VI-2). The Q-values estimated using the eight SNP assays, or their combinations, were compared with those obtained from the whole-genome dataset (2.336 M SNPs), which is assumed to provide the true admixture proportions. The Q-values obtained with M1, M2, M3, and M4 are highly correlated with those of the whole-genome dataset ($0.956 < r < 0.982$; Table VI-4, Figure Sup VI-4). While all statistics indicate that the four reduced assays have a good performance, M2 shows consistently the worst behaviour. The mean accuracy, for example, is high across the assays, varying between 95.93% (M2) and 97.42% (M1), but the dispersion is much greater for M2 (Table VI-5, Figure VI-4).

Interestingly, the four random SNP assays also show a good performance, although M3 and M4 are considerably better, as indicated by the non-overlapping confidence intervals of the correlations (Table VI-5, Figure Sup VI-4) and the lower dispersion of the accuracy values around the median (Figure VI-4). Another important difference between M and R assays arises from the misclassification of individuals and simulated haplotypes (pure classified as hybrid and *vice-versa*), with the reduced assays performing consistently better than the random ones. For example, all

random assays misclassified between one to three pure individuals as hybrids, which never occurred with the reduced assays (Table VI-5, Table Sup VI-8).

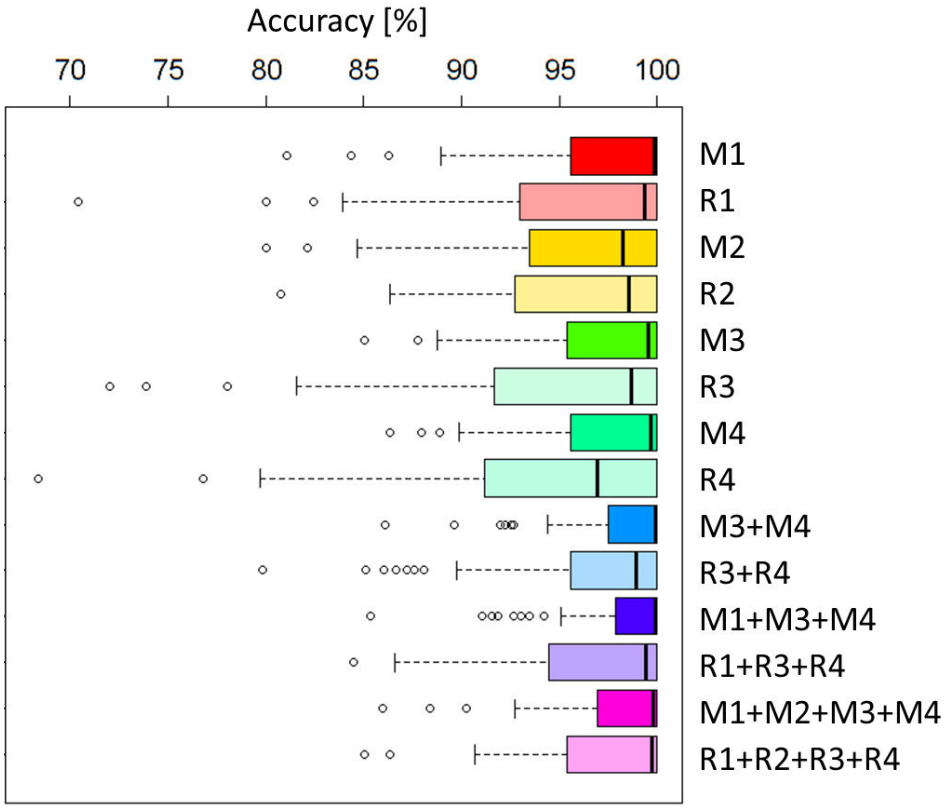


Figure VI-4 - Accuracy of single and combined reduced (M1-M4) and random (R1-R4) SNP assays. The box denotes the first and third quartiles and median accuracy marked with a bold vertical line within the box. Outliers are indicated by circles. Random assays consistently have a larger inter-quartile range than the corresponding reduced assay

The overall performance increases when the reduced assays are combined (Table VI-5, S8; Figure VI-4, Sup VI-4). The best result is obtained for the combination of M1, M3 and M4, which represents a total of 115 highly informative SNPs distributed across the 16 chromosomes. However, the combination of M3 and M4, with only 78 SNPs, was nearly as good (Table VI-5). In summary, while there is an increment in the overall performance when combining M1, M3, and M4, their individual use still provides robust estimates of C-lineage introgression into *A. m. iberiensis*.

Table VI-5 - Performance of the reduced (M1-M4) and random (R1-R4) SNP assays in estimating C-lineage introgression (Q-values) of holdout and simulated datasets as compared to the whole-genome dataset. *(i)* Pearson's correlation coefficient r; *(ii)* mean standard error estimated from 200 bootstrap replicates by ADMIXTURE; *(iii)* mean error calculated by the absolute difference; *(iv)* number of individuals with error >0.05; *(v)* maximum error; *(vi)* mean accuracy calculated via percentage of absolute error; *(vii)* precision defined as the standard deviation of the absolute error; *(viii)* number of misclassified individuals (Q-value threshold of 0.05)

Assay	# of SNPs	Pearson's r (95% CI) (*i*)	Standard error (*ii*)	Mean error (*iii*)	# Ind error 0.05 (*iv*)	Max error (*v*)	% Mean accuracy (*vi*)	Precision (*vii*)	Pure classified as hybrid (*viii*)	Hybrid classified as pure
M1	37	0.975 (0.958-0.985)	0.046	0.026	12	0.189	97.42	0.043	0	0
R1	37	0.949 (0.915-0.970)	0.069	0.043	20	0.296	95.71	0.062	1	3
M2	38	0.956 (0.927-0.974)	0.046	0.041	20	0.200	95.93	0.053	1	0
R2	38	0.967 (0.945-0.981)	0.075	0.037	20	0.192	96.34	0.047	3	1
M3	40	0.978 (0.964-0.987)	0.048	0.028	13	0.15	97.24	0.038	0	0
R3	40	0.933 (0.888-0.960)	0.067	0.05	14	0.279	95.04	0.069	1	1
M4	38	0.982 (0.969-0.989)	0.044	0.026	13	0.137	97.41	0.036	1	0
R4	38	0.925 (0.876-0.955)	0.062	0.053	22	0.316	94.71	0.069	3	1
M3+M4	78	0.988 (0.979-0.993)	0.04	0.018	9	0.139	98.18	0.030	0	0
R3+R4	78	0.967 (0.945-0.981)	0.051	0.034	13	0.201	96.62	0.049	1	0
M1+M3+M4	115	0.987 (0.979-0.993)	0.037	0.018	8	0.147	98.15	0.030	0	0
R1+R3+R4	115	0.976 (0.959-0.986)	0.046	0.03	16	0.155	97.01	0.041	0	1
M1+M2+M3+M4	153	0.986 (0.977-0.992)	0.003	0.02	9	0.14	98.02	0.031	0	0
R1+R2+R3+R4	153	0.981 (0.967-0.989)	0.042	0.027	14	0.15	97.35	0.037	0	1

Discussion

Developing cost-effective molecular tools for accurate estimation of introgression in *A. mellifera* is increasingly important as commercial strains (mostly of C-lineage ancestry) are threatening native genetic diversity in many regions throughout Europe (Bertrand *et al.*, 2015; De la Rúa *et al.*, 2009; Jensen *et al.*, 2005; Parejo *et al.*, 2016; Pinto *et al.*, 2014; Soland-Reckeweg *et al.*, 2009). In the post-genomics era, rapid innovations in high-throughput sequencing technologies make it possible to construct extensive whole-genome datasets, especially in model organisms with small genomes like the honey bee (Weinstock *et al.*, 2006). However, while whole-genome sequencing is increasingly inexpensive (~200 €/honey bee), it is still not affordable for conservation management applications. Furthermore, the processing of the large amounts of data generated by whole-genome sequencing requires bioinformatics expertise and powerful computational resources typically not available to state entities or conservation centres. Whole-genome sequences, however, can be used to generate baseline data for developing robust molecular tools for routine genotyping hundreds of samples in a time and cost-effective manner. Here, we mined a massive whole-genome dataset, representing the focal *A. m. iberiensis* and the two C-lineage subspecies (*A. m. carnica* and *A. m. ligustica*) preferred worldwide in commercial breeding, to identify fixed SNPs for constructing robust reduced assays. This approach represents a rigorous methodological example that can be applied for developing reduced SNP assays in any other organism.

Effect of sampling bias on the number of fixed SNPs

Considering the long-standing problem of ascertainment bias during discovery and selection of informative SNPs (Albrechtsen *et al.*, 2010, and references therein), we started by testing the effect of sample size and sampling breadth on the number of SNPs erroneously identified as fixed between *A. m. iberiensis* and C-lineage (false positive fixed SNPs). We found that limited sample size can be problematic, as a considerable number of false positive fixed SNPs with F_{ST} ≤0.95 could negatively impact the development of a sensitive SNP assay. This effect is reduced for N=25, and increasing sample size above 50 yields diminishing returns in fixed SNPs, suggesting that an optimal cost-benefit ratio is reached. Beyond this point, further increasing sample size will likely lead to detection of new SNPs in the population. However, such low-frequency SNPs (i.e. singletons) are not of concern for discriminating populations, nor for identifying highly informative SNPs.

A bias is also introduced when sampling a geographically restricted area. From the three geographic subsets examined, the Portuguese revealed the highest number of false positives while the Central and Mediterranean behaved similarly to the subset covering the entire Iberian honey bee range. While both the Central and Mediterranean subsets cover the northeastern-southwestern Iberian cline, the Portuguese subset represents a small portion of the *A. m. iberiensis* genetic complexity (Chávez-Galarza *et al.*, 2017; Chávez-Galarza *et al.*, 2015; Pinto *et al.*, 2013). But more importantly, this subset generated a substantial number of false positives with a lower differentiation power (Table VI-4). As a consequence, reduced SNP assays designed from samples strictly originating from Portugal would not be appropriate to discriminate *A. m. iberiensis* from C-lineage, but only the Portuguese populations. While selecting informative SNPs from geographically-limited samples or subpopulations may be valid for very specific applications, it is not a recommended procedure in most cases (especially when knowledge on population structure is lacking) and questions the wider applicability of SNP assays. It is well established, that this kind of ascertainment bias influences population genetic measures such as divergence (Albrechtsen *et al.*, 2010) and demography (Morin *et al.*, 2004; Wakeley *et al.*, 2001). Accordingly, we assured a sufficiently large and representative sample of the *A. m. iberiensis* diversity, which covers the Iberian cline, for developing accurate reduced assays while at the same time leaving independent holdout samples for validation.

Genomic information of the highly informative SNPs

A large number of SNPs (18,272) were fixed between *A. m. iberiensis* and C-lineage subspecies. This was an expected result because M and C are the most divergent of the four lineages (Wallberg *et al.*, 2014). The top enriched GO terms of the genes marked by those SNPs were associated with numerous genes related to regulation of expression, which is essential for the versatility and adaptability of a species for short- and long-term environmental changes (López-Maury *et al.*, 2008). This is consistent with the complex evolutionary history of *A. mellifera*, and its numerous subspecies, which has adapted to the diversity of habitats and climates in its large distributional range (Harpur *et al.*, 2014; Wallberg *et al.*, 2014).

Assay design and validation

Having a large number of fixed SNPs is an enormous advantage when designing reduced SNP assays, as they represent ideal ancestry informative markers (Rosenberg *et al.*, 2003). Yet, the overall high differentiation between *A. m. iberiensis* and C-lineage honey bees explains why all tested assays, including those constructed from randomly selected SNPs, performed well. For example, a random set of 153 SNPs performed equally well as the 153 fixed SNPs across the four reduced assays. This was also shown by Pardo-Seco *et al.* (2014) who concluded that it is not primarily individual informativeness, but the number of markers that plays a major role in accurately estimating genome ancestry. Although all the assays show a remarkable performance on average, we highlight, however, that differences arise at the individual level. While average statistics can be useful for measuring the admixture proportions of an entire population, they are not adequate to support decision making at the individual level, e.g. when choosing individuals for conservation breeding purposes. Three random assays had individual errors >25% compared to the whole-genome information, which is far from acceptable in a context of conservation. Moreover, pure *A. m. iberiensis*, which were misclassified as hybrids, could lead to exclusion of individuals with valuable and unique genetic components.

Apart from assay performance, the genotyping cost is another important criterion to take into consideration. Genotyping with the MassARRAY® system costs approximately 5.5€ per individual and single assay. While the M1, M3, and M4 perform remarkably well, the minimal individual error and the highest accuracy is achieved when combining the three assays (115 SNPs), although the combination of M3 and M4 (78 SNPs) is nearly as good. The choice of using up to three assays is ultimately dictated by budget constraints; nevertheless, an interesting trade-off between accuracy and cost is achieved when genotyping the 78 SNPs.

Unlike many populations of *A. m. mellifera* from Western Europe and *A.m. iberiensis* from the archipelagos of Baleares and Macaronesia, which are threatened by human-mediated gene flow (De La Rúa *et al.*, 2001; De la Rúa *et al.*, 2003; Jensen *et al.*, 2005; Miguel *et al.*, 2015; Muñoz *et al.*, 2014; Pinto *et al.*, 2014), there is very limited introgression in *A. m. iberiensis* populations of Iberia (Chávez-Galarza *et al.*, 2015). Therefore, it is crucial to monitor Iberian populations, before gene complexes shaped by natural selection over evolutionary time are irretrievably lost. Here, we took advantage of whole-genome sequence data, which provided millions of SNPs, to design highly powerful assays containing a low number of SNPs capable of estimating C-lineage introgression

into *A. m. iberiensis* with a high level of accuracy. We recommend the combination of the best two (78 SNPs) or three (115 SNPs) reduced SNP assays, although one assay can also be used when there are budget constraints. These assays can be used to estimate C-lineage introgression not only in the native range of *A. m. iberiensis* in Iberia but also in the introduced range in the archipelagos of Baleares and Macaronesia, and in South America.

This study provides a powerful set of tools to safeguard a unique legacy of honey bee diversity for future generations. While these tools can only be applied to honey bees, the approach demonstrated herein (from testing the effect of sampling bias to the intricate steps involved in the design of the reduced SNP assays) is of high general value in a wide range of scenarios for the conservation of potentially hybridized domestic and wildlife populations.

Data Accessibility

Sequence data of *A. m. iberiensis* is available at: to be completed after manuscript is accepted for publication. *A. m. carnica* sequence data is deposited at the ENA (www.ebi.ac.uk/ena) under study accession number PRJEB16533 and *A. m. ligustica* sequences are available through the SeqApiPop programme on SRA (www.ncbi.nlm.nih.gov/sra) with accession number PRJNA311274.

Acknowledgements

We thank numerous researchers, beekeepers and beekeeping associations who provided samples and assisted with sampling. João Costa, Instituto Gulbenkian Ciência, designed the multiplexes with the Assay Design tool. José Rufino provided computational resources at IPB. John C. Patton, Phillip San Miguel, Paul Parker, Rick Westerman, University of Purdue, sequenced most honey bees and many reference samples. Reference samples were also sequenced at the GeT PlaGe platform in Toulouse. An earlier version of the manuscript was improved by the constructive comments made by two anonymous reviewers. Dora Henriques was supported by a PhD scholarship from the Fundação para a Ciência e Tecnologia (FCT) (SFRH/BD/84195/2012) and Melanie Parejo by the Swiss Federal Office for Agriculture FOAG and the Sur-la-Croix foundation, Basel. Analyses were performed at UPPMAX, Uppsala University, and UBELIX, University of Bern. MAP is a member of and receives support from the COST Action FA1307 (SUPER-B). This research was funded through the projects PTDC/BIA-BEC/099640/2008 (FCT and COMPETE/QREN/EU) and the 2013-2014 BiodivERsA/FACCE-JPI joint call for research proposals, with the national

funders FCT (Portugal), "Agence Nationale de la Recherche" (France), and "Ministério de Economia y Competividade" (Spain).

References

Albrechtsen, A., Nielsen, F. C., & Nielsen, R. (2010). Ascertainment biases in SNP chips affect measures of population divergence. *Molecular Biology and Evolution, 27*(11), 2534-2547. doi: 10.1093/molbev/msq148

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Resources*. doi: 10.1101/gr.094052.109

Amirisetty, S., Khurana Hershey, G. K., & Baye, T. M. (2012). AncestrySNPminer: A bioinformatics tool to retrieve and develop ancestry informative SNP panels. *Genomics, 100*(1), 57-63. doi: 10.1016/j.ygeno.2012.05.003

Anderson, E. C. (2010). Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. *Molecular Ecology Resources, 10*(4), 701-710. doi: 10.1111/j.1755-0998.2010.02846.x

Arias, M. C., Rinderer, T. E., & Sheppard, W. S. (2006). Further characterization of honey bees from the Iberian Peninsula by allozyme, morphometric and mtDNA haplotype analyses. *Journal of Apicultural Research, 45*(4), 188-196. doi: 10.1080/00218839.2006.11101346

Bertrand, B., Alburaki, M., Legout, H., Moulin, S., Mougel, F., & Garnery, L. (2015). MtDNA COI-COII marker and drone congregation area: An efficient method to establish and monitor honeybee (*Apis mellifera* L.) conservation centres. *Molecular Ecology Resources, 15*(3), 673-683. doi: 10.1111/1755-0998.12339

Büchler, R., Costa, C., Hatjina, F., Andonov, S., Meixner, M. D., Conte, Y. L., Uzunov, A., Berg, S., Bienkowska, M., & Bouga, M. (2014). The influence of genetic origin and its interaction with environmental effects on the survival of *Apis mellifera* L. colonies in Europe. *Journal of Apicultural Research, 53*(2), 205-214. doi: 10.3896/IBRA.1.53.2.03

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience, 4*, 7. doi: 10.1186/s13742-015-0047-8

Chapman, N. C., Harpur, B. A., Lim, J., Rinderer, T. E., Allsopp, M. H., Zayed, A., & Oldroyd, B. P. (2015). A SNP test to identify Africanized honeybees via proportion of 'African' ancestry. *Molecular Ecology Resources, 15*(6), 1346-1355. doi: 10.1111/1755-0998.12411

Chávez-Galarza, J., Garnery, L., Henriques, D., Neves, C. J., Loucif-Ayad, W., Jonhston, J. S., & Pinto, M. A. (2017). Mitochondrial DNA variation of *Apis mellifera iberiensis*: further insights from a large-scale

study using sequence data of the tRNAleu-cox2 intergenic region. *Apidologie*, 1-12. doi: 10.1007/s13592-017-0498-2

Chávez-Galarza, J., Henriques, D., Johnston, J. S., Azevedo, J. C., Patton, J. C., Muñoz, I., De la Rúa, P., & Pinto, M. A. (2013). Signatures of selection in the Iberian honey bee (*Apis mellifera iberiensis*) revealed by a genome scan analysis of single nucleotide polymorphisms. *Molecular Ecology, 22*(23), 5890-5907. doi: 10.1111/mec.12537

Chávez-Galarza, J., Henriques, D., Johnston, J. S., Carneiro, M., Rufino, J., Patton, J. C., & Pinto, M. A. (2015). Revisiting the Iberian honey bee (*Apis mellifera iberiensis*) contact zone: maternal and genome-wide nuclear variations provide support for secondary contact from historical refugia. *Molecular Ecology, 24*(12), 2973-2992. doi: 10.1111/mec.13223

Chen, C., Liu, Z., Pan, Q., Chen, X., Wang, H., Guo, H., Liu, S., Lu, H., Tian, S., Li, R., & Shi, W. (2016). Genomic Analyses Reveal Demographic History and Temperate Adaptation of the Newly Discovered Honey Bee Subspecies *Apis mellifera sinisxinyuan n. ssp*. *Molecular Biology and Evolution, 33*(5), 1337-1348. doi: 10.1093/molbev/msw017

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly, 6*(2), 80-92. doi: 10.4161/fly.19695

De La Rúa, P., Galián, J. A., Serrano, J., & Moritz, R. F. A. (2001). Genetic structure and distinctness of *Apis mellifera* L. populations from the Canary Islands. *Molecular Ecology, 10*(7), 1733-1742. doi: 10.1046/j.1365-294X.2001.01303.x

De la Rúa, P., Galián, J. A., Serrano, J., & Moritz, R. F. A. (2003). Genetic structure of Balearic honeybee populations based on microsatellite polymorphism. *Genetics, Selection, Evolution : GSE, 35*(3), 339-350. doi: 10.1186/1297-9686-35-3-339

De la Rúa, P., Jaffé, R., Dall'Olio, R., Muñoz, I., & Serrano, J. (2009). Biodiversity, conservation and current threats to European honeybees. *Apidologie, 40*(3), 263-284. doi: 10.1051/apido/2009027

De la Rúa, P., Jaffé, R., Muñoz, I., Serrano, J., Moritz, R. F. A., & Kraus, F. B. (2013). Conserving genetic diversity in the honeybee: Comments on Harpur et al.(2012). *Molecular Ecology, 22*(12), 3208-3210. doi: 10.1111/mec.12333

Ding, L., Wiener, H., Abebe, T., Altaye, M., Go, R. C., Kercsmar, C., Grabowski, G., Martin, L. J., Khurana Hershey, G. K., Chakorborty, R., & Baye, T. M. (2011). Comparison of measures of marker informativeness for ancestry and admixture mapping. *BMC Genomics, 12*(1), 622. doi: 10.1186/1471-2164-12-622

Engel, M. S. (1999). The taxonomy of recent and fossil honey bees (Hymenoptera: Apidae; Apis). *J. Hym. Res, 8*(2).

Francis, R. M., Amiri, E., Meixner, M. D., Kryger, P., Gajda, A., Andonov, S., Uzunov, A., Topolska, G., Charistos, L., Costa, C., Berg, S., Bienkowska, M., Bouga, M., Büchler, R., Dyrba, W., Hatjina, F., Ivanova, E., Kezic, N., Korpela, S., Conte, Y. L., Panasiuk, B., Pechhacker, H., Tsoktouridis, G., & Wilde, J. (2014). Effect of genotype and environment on parasite and pathogen levels in one apiary—a case study. *Journal of Apicultural Research, 53*(2), 230-232. doi: 10.3896/IBRA.1.53.2.14

Franck, P., Garnery, L., Solignac, M., & Cornuet, J.-M. (1998). The Origin of West European Subspecies of Honeybees (*Apis mellifera*): New Insights from Microsatellite and Mitochondrial Data. *Evolution, 52*(4), 1119-1134. doi: 10.2307/2411242

Frankham, R., Ballou, J. D., & Briscoe, D. A. (2002). *Introduction to Conservation Genetics*: Cambridge University Press.

Gene Ontology Consortium. (2015). Gene ontology consortium: going forward. *Nucleic acids research, 43*(D1), D1049-D1056. doi: 10.1093/nar/gku1179

Harpur, B. A., Kent, C. F., Molodtsova, D., Lebon, J. M., Alqarni, A. S., Owayss, A. A., & Zayed, A. (2014). Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. *Proc Natl Acad Sci U S A, 111*(7), 2614-2619. doi: 10.1073/pnas.1315506111

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res, 37*(1), 1-13. doi: 10.1093/nar/gkn923

Hulsegge, B., Calus, M., Windig, J., Hoving-Bolink, A., Maurice-van Eijndhoven, M., & Hiemstra, S. (2013). Selection of SNP from 50K and 777K arrays to predict breed of origin in cattle. *Journal of Animal Science, 91*(11), 5128-5134. doi: 10.2527/jas.2013-6678

Jensen, A. B., Palmer, K. A., Boomsma, J. J., & Pedersen, B. V. (2005). Varying degrees of *Apis mellifera ligustica* introgression in protected populations of the black honeybee, *Apis mellifera mellifera*, in northwest Europe. *Molecular Ecology, 14*(1), 93-106. doi: 10.1111/j.1365-294X.2004.02399.x

Judge, M., Kelleher, M., Kearney, J., Sleator, R., & Berry, D. (2017). Ultra-low-density genotype panels for breed assignment of Angus and Hereford cattle. *Animal, 11*(6), 938-947. doi: 10.1017/S1751731116002457

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res, 44*(Database issue), D457-D462. doi: 10.1093/nar/gkv1070

Karlsson, S., Moen, T., Lien, S., Glover, K. A., & Hindar, K. (2011). Generic genetic differences between farmed and wild Atlantic salmon identified from a 7K SNP-chip. *Molecular Ecology Resources, 11*(s1), 247-253. doi: 10.1111/j.1755-0998.2010.02959.x

Le Conte, Y., & Navajas, M. (2008). Climate change: impact on honey bee populations and diseases. *Revue Scientifique et Technique-Office International des Epizooties, 27*(2), 499-510.

López-Maury, L., Marguerat, S., & Bahler, J. (2008). Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nat Rev Genet, 9*(8), 583-593. doi: 10.1038/nrg2398

Mariette, S., Le Corre, V., Austerlitz, F., & Kremer, A. (2002). Sampling within the genome for measuring within-population diversity: trade-offs between markers. *Molecular Ecology, 11*(7), 1145-1156. doi: 10.1046/j.1365-294X.2002.01519.x

Meixner, M. D., Costa, C., Kryger, P., Hatjina, F., Bouga, M., Ivanova, E., & Büchler, R. (2010). Conserving diversity and vitality for honey bee breeding. *Journal of Apicultural Research, 49*(1), 85-92. doi: 10.3896./IBRA.1.49.1.12

Meixner, M. D., Leta, M. A., Koeniger, N., & Fuchs, S. (2011). The honey bees of Ethiopia represent a new subspecies of *Apis mellifera- Apis mellifera simensis* n. ssp. . *Apidologie, 42*, 425-437. doi: 10.1007/s13592-011-0007-y

Miguel, I., Garnery, L., Iriondo, M., Baylac, M., Manzano, C., Steve Sheppard, W., & Estonba, A. (2015). Origin, evolution and conservation of the honey bees from La Palma Island (Canary Islands): molecular and morphological data. *Journal of Apicultural Research, 54*(5), 427-440. doi: 10.1080/00218839.2016.1180017

Miguel, I., Iriondo, M., Garnery, L., Sheppard, W. S., & Estonba, A. (2007). Gene flow within the M evolutionary lineage of *Apis mellifera*: role of the Pyrenees, isolation by distance and post-glacial re-colonization routes in the western Europe. *Apidologie, 38*(2), 141-155. doi: 10.1051/apido:2007007

Morin, P. A., Luikart, G., Wayne, R. K., & Grp, S. N. P. W. (2004). SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution, 19*. doi: 10.1016/j.tree.2004.01.009

Muñoz, I., Henriques, D., Jara, L., Johnston, J. S., Chávez-Galarza, J., De La Rúa, P., & Pinto, M. A. (2017). SNPs selected by information content outperform randomly selected microsatellite loci for delineating genetic identification and introgression in the endangered Dark European honeybee (*Apis mellifera mellifera*). *Molecular Ecology Resources, 17*(4), 783-795. doi: 10.1111/1755-0998.12637

Muñoz, I., Henriques, D., Johnston, J. S., Chávez-Galarza, J., Kryger, P., & Pinto, M. A. (2015). Reduced SNP Panels for Genetic Identification and Introgression Analysis in the Dark Honey Bee (*Apis mellifera mellifera*). *PLoS One, 10*(4), e0124365. doi: 10.1371/journal.pone.0124365

Muñoz, I., Pinto, M. A., & De la Rúa, P. (2014). Effects of queen importation on the genetic diversity of Macaronesian island honey bee populations (*Apis mellifera* Linneaus 1758). *Journal of Apicultural Research, 53*(2), 296-302. doi: 10.3896/IBRA.1.53.2.11

Neumann, P., & Blacquière, T. (2017). The Darwin cure for apiculture? Natural selection and managed honeybee health. *Evolutionary Applications, 10*(3), 226-230. doi: 10.1111/eva.12448

Pardo-Seco, J., Martinón-Torres, F., & Salas, A. (2014). Evaluating the accuracy of AIM panels at quantifying genome ancestry. *BMC Genomics, 15*(1), 543. doi: 10.1186/1471-2164-15-543

Parejo, M., Wragg, D., Gauthier, L., Vignal, A., Neumann, P., & Neuditschko, M. (2016). Using Whole-genome Sequence Information to Foster Conservation Efforts for the European Dark Honey Bee, *Apis mellifera mellifera*. *Frontiers in Ecology and Evolution, 4*(140). doi: 10.3389/fevo.2016.00140

Pinto, M. A., Henriques, D., Chávez-Galarza, J., Kryger, P., Garnery, L., van der Zee, R., Dahle, B., Soland-Reckeweg, G., De la Rúa, P., Dall' Olio, R., Carreck, N., & Johnston, J. S. (2014). Genetic integrity of the Dark European honey bee (*Apis mellifera mellifera*) from protected populations: a genome-wide assessment using SNPs and mtDNA sequence data. *Journal of Apicultural Research, 53*(2), 269-278. doi: 10.3896/ibra.1.53.2.08

Pinto, M. A., Henriques, D., Guedes, H., Munoz, I., Azevedo, J. C., & De la Rúa, P. (2013). Maternal diversity patterns of Ibero-Atlantic populations reveal further complexity of Iberian honeybees. *Apidologie 44* 430-439. doi: 10.1007/s13592-013-0192-y

Potts, S. G., Biesmeijer, J. C., Kremen, C., Neumann, P., Schweiger, O., & Kunin, W. E. (2010). Global pollinator declines: trends, impacts and drivers. *Trends in Ecology & Evolution, 25*(6), 345-353. doi: 10.1016/j.tree.2010.01.007

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., & McGarvey, K. M. (2013). RefSeq: an update on mammalian reference sequences. *Nucleic acids research, 42*(D1), D756-D763. doi: 10.1093/nar/gkt1114

R Core Team (Ed.). (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rosenberg, N. A., Li, L. M., Ward, R., & Pritchard, J. K. (2003). Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet, 73*. doi: 10.1086/380416

Ruttner, F. (1988). *Biogeography and Taxonomy of Honey Bees*: Springer, Berlin.

Sheppard, W. S., & Meixner, M. D. (2003). *Apis mellifera pomonella*, a new honey bee subspecies from Central Asia. *Apidologie, 34*, 367–375. doi: 10.1051/apido:2003037

Smith, D. R., Palopoli, M. F., Taylor, B. R., Garnery, L., Cornuet, J. M., Solignac, M., & Brown, W. M. (1991). Geographical overlap of two mitochondrial genomes in Spanish honeybees (*Apis mellifera iberica*). *Journal of Heredity, 82*(2), 96-100.

Soland-Reckeweg, G., Heckel, G., Neumann, P., Fluri, P., & Excoffier, L. (2009). Gene flow in admixed populations and implications for the conservation of the Western honeybee, *Apis mellifera*. *Journal of Insect Conservation, 13*(3), 317-328. doi: 10.1007/s10841-008-9175-0

vanEngelsdorp, D., & Meixner, M. D. (2010). A historical review of managed honey bee populations in Europe and the United States and the factors that may affect them. *Journal of Invertebrate Pathology, 103, Supplement*, S80-S95. doi: 10.1016/j.jip.2009.06.011

Vignal, A., Milan, D., SanCristobal, M., & Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics, Selection, Evolution : GSE, 34*(3), 275-305. doi: 10.1186/1297-9686-34-3-275

Wakeley, J., Nielsen, R., Liu-Cordero, S. N., & Ardlie, K. (2001). The discovery of single nucleotide polymorphisms and inferences about human demographic history. *The American Journal of Human Genetics, 69*, 1332-1347. doi: 10.1086/324521

Wallberg, A., Glémin, S., & Webster, M. T. (2015). Extreme recombination frequencies shape genome variation and evolution in the honeybee, *Apis mellifera*. *PLoS genetics, 11*(4). doi: 10.1371/journal.pgen.1005189

Wallberg, A., Han, F., Wellhagen, G., Dahle, B., Kawata, M., Haddad, N., Simões, Z. L. P., Allsopp, M. H., Kandemir, I., De la Rúa, P., Pirk, C. W., & Webster, M. T. (2014). A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee *Apis mellifera*. *Nature Genetics, 46*, 1081. doi: 10.1038/ng.3077

Weinstock, G. M., Robinson, G. E., Gibbs, R. A., Worley, K. C., Evans, J. D., Maleszka, R., Robertson, H. M., Weaver, D. B., Beye, M., Bork, P., Elsik, C. G., Hartfelder, K., Hunt, G. J., Zdobnov, E. M., Amdam, G. V., Bitondi, M. M. G., Collins, A. M., Cristino, A. S., Lattorff, H. M. G., Lobo, C. H., Moritz, R. F. A., Nunes, F. M. F., Page, J. R. E., Simoes, Z. L. P., Wheeler, D., Carninci, P., Fukuda, S., Hayashizaki, Y., Kai, C., Kawai, J., Sakazume, N., Sasaki, D., Tagami, M., Albert, S., Baggerman, G., Beggs, K. T., Bloch, G., Cazzamali, G., Cohen, M., Drapeau, M. D., Eisenhardt, D., Emore, C., Ewing, M. A., Fahrbach, S. E., Foret, S., Grimmelikhuijzen, C. J. P., Hauser, F., Hummon, A. B., Huybrechts, J., Jones, A. K., Kadowaki, T., Kaplan, N., Kucharski, R., Leboulle, G., Linial, M., Littleton, J. T., Mercer, A. R., Richmond, T. A., Rodriguez-Zas, S. L., Rubin, E. B., Sattelle, D. B., Schlipalius, D., Schoofs, L., Shemesh, Y., Sweedler, J. V., Velarde, R., Verleyen, P., Vierstraete, E.,

Williamson, M. R., Ament, S. A., Brown, S. J., Corona, M., Dearden, P. K., Dunn, W. A., Elekonich, M. M., Fujiyuki, T., Gattermeier, I., Gempe, T., Hasselmann, M., Kadowaki, T., Kage, E., Kamikouchi, A., Kubo, T., Kucharski, R., Kunieda, T., Lorenzen, M. D., Milshina, N. V., Morioka, M., Ohashi, K., Overbeek, R., Ross, C. A., Schioett, M., Shippy, T., Takeuchi, H., Toth, A. L., Willis, J. H., Wilson, M. J., Gordon, K. H. J., Letunic, I., Hackett, K., Peterson, J., Felsenfeld, A., Guyer, M., Solignac, M., Agarwala, R., Cornuet, J. M., Monnerot, M., Mougel, F., Reese, J. T., Vautrin, D., Gillespie, J. J., Cannone, J. J., Gutell, R. R., Johnston, J. S., Eisen, M. B., Iyer, V. N., Iyer, V., Kosarev, P., Mackey, A. J., Solovyev, V., Souvorov, A., Aronstein, K. A., Bilikova, K., Chen, Y. P., Clark, A. G., Decanini, L. I., Gelbart, W. M., Hetru, C., Hultmark, D., Imler, J.-L., Jiang, H., Kanost, M., Kimura, K., Lazzaro, B. P., Lopez, D. L., Simuth, J., Thompson, G. J., Zou, Z., De Jong, P., Sodergren, E., Csuroes, M., Milosavljevic, A., Osoegawa, K., Richards, S., Shu, C.-L., Duret, L., Elhaik, E., Graur, D., Anzola, J. M., Campbell, K. S., Childs, K. L., Collinge, D., Crosby, M. A., Dickens, C. M., Grametes, L. S., Grozinger, C. M., Jones, P. L., Jorda, M., Ling, X., Matthews, B. B., Miller, J., Mizzen, C., Peinado, M. A., Reid, J. G., Russo, S. M., Schroeder, A. J., St Pierre, S. E., Wang, Y., Zhou, P., Jiang, H., Kitts, P., Ruef, B., Venkatraman, A., Zhang, L., Aquino-Perez, G., Whitfield, C. W., Behura, S. K., Berlocher, S. H., Sheppard, W. S., Smith, D. R., Suarez, A. V., Tsutsui, N. D., Wei, X., Wheeler, D., Havlak, P., Li, B., Liu, Y., Sodergren, E., Jolivet, A., Lee, S., Nazareth, L. V., Pu, L.-L., Thorn, R., Stolc, V., Newman, T., Samanta, M., Tongprasit, W. A., Claudianos, C., Berenbaum, M. R., Biswas, S., de Graaf, D. C., Feyereisen, R., Johnson, R. M., Oakeshott, J. G., Ranson, H., Schuler, M. A., Muzny, D., Chacko, J., Davis, C., Dinh, H., Gill, R., Hernandez, J., Hines, S., Hume, J., Jackson, L., Kovar, C., Lewis, L., Miner, G., Morgan, M., Nguyen, N., Okwuonu, G., Paul, H., Santibanez, J., Savery, G., Svatek, A., Villasana, D., & Wright, R. (2006). Insights into social insects from the genome of the honeybee *Apis mellifera*. *NATURE, 443*, 931-949.

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution, 38*. doi: 10.2307/2408641

Wilkinson, S., Wiener, P., Archibald, A. L., Law, A., Schnabel, R. D., McKay, S. D., Taylor, J. F., & Ogden, R. (2011). Evaluation of approaches for identifying population informative markers from high density SNP Chips. *BMC Genetics, 12*(1), 45. doi: 10.1186/1471-2156-12-45

Chapter VII.

Mitochondrial DNA patterns in honey bees: from mitogenomes to the popular intergenic $tRNA^{leu}$-cox2 region

Henriques D, Chávez-Galarza J, Neves C., Quaresma A., Costa F., Rufino J, Pinto MA.

Abstract

The maternal honey bee genetic variation has been thoroughly assessed using the highly polymorphic tRNA^{leu}-cox2 intergenic region. However, different mitogenome regions evolve at different rates and different genes may lead to incongruent results, so it is important to understand the information content of each single gene. Using a mtDNA-WGS dataset of 123 individuals representing seven subspecies, three lineages (A, M and C) and three African sub-lineages (A_I, A_{II} and A_{III}), we studied the mitogenomic phylogenies and phylogeography through the analysis of both individual genes and the complete mitogenome, in order to compare this information with that obtained from the tRNA^{leu}-cox2 intergenic region. We show that the popular tRNA^{leu}-cox2 intergenic region does not parallel the evolutionary history of the mitogenome. While the mitogenome and the single genes support the three evolutionary lineages defined by the *Dra*I restriction pattern of the tRNA^{leu}-cox2 intergenic region, the African sub-lineages and even the haplotypes are not supported. Therefore, it is important to understand that depending on the aim of the work the tRNA^{leu}-cox2 intergenic region may not be appropriate to study relationships between individuals.

Keywords: Honey bee, tRNA^{leu}-cox2, mitogenome

Introduction

Mitochondrial DNA (mtDNA) is one of the most popular molecular markers to assess genetic diversity and phylogeographical patterns. The popularity of mtDNA as a marker can be explained by practical reasons. First, mtDNA is easy to amplify using ''universal'' polymerase chain reaction (PCR) primers since it is very abundant in the cells. Second, its non-recombining maternal inheritance makes the phylogenetic and phylogeographical interpretation straightforward. Finally, it has a rapid mutation rate and short coalescence time that makes this marker suitable to study recent population genetic events (Boore *et al.*, 2005; Brito & Edwards, 2009; Ma *et al.*, 2012; Meixner *et al.*, 2013).

Since seminal work of Avise and Ellis (1986) supporting the usefulness of mtDNA markers to phylogeographic inference, hundreds of studies have been performed with short segments of mtDNA. These short segments of mtDNA can provide an overview of phylogeographical patterns. However, they may have limited resolution to solve very recent events (Jacobsen *et al.*, 2012). The advent of second generation sequencing technologies turned the sequencing of the whole mitochondrial genome (mitogenome) fast and economical (Jacobsen *et al.*, 2012). The mitogenome has higher resolution than single genes and has been applied to a variety of phylogenetic and phylogeographic studies to solve shallow evolutionary histories mainly in a recent timescale, such as in the brown bear (Keis *et al.*, 2013), wolly mammoth (Gilbert *et al.*, 2008), dog (Pang *et al.*, 2009), cattle (Achilli *et al.*, 2008), yak (Wang *et al.*, 2010), killer whale (Morin *et al.*, 2010), giant squid (Winkelmann *et al.*, 2013), speartooth shark (Feutry *et al.*, 2017; Feutry *et al.*, 2014), bank vole (Filipi *et al.*, 2015), and snub-nosed monkey (Hong *et al.*, 2017).

In 1985, the *Drosophila yakuba* mitogenome was first published (Clary & Wolstenholme 1985) and since then, other insect mitogenomes have been sequenced, including that of the Italian honey bee (*Apis mellifera ligustica*), which was published in 1993 (Crozier & Crozier, 1993). Such as in most of the Metazoa, the mitochondrial honey bee genome contains 37 genes: two rDNAs that encode the mitochondrial ribosome RNA components; 22 tRNAs required for translation of the mitochondrial proteins and 13 protein-coding genes encoding the essential components of the electron transport chain and oxidative phosphorylation (Arif & Khan, 2009).

Mitochondrial DNA has been widely applied to study the genetic diversity and population structure of honey bees, using a range of molecular methods, such as RFLPs (Restriction Fragment Length Polymorphisms; (Garnery *et al.*, 1992; Hall & Smith, 1991) PCR-RFLPs (Chávez-Galarza *et*

al., 2015; Pinto *et al.*, 2004) and direct sequencing (Arias & Sheppard, 1996; Chávez-Galarza *et al.*, 2017; Techer *et al.*, 2017). Initially, the whole mitogenome was surveyed with the RFLP method (Garnery *et al.*, 1992; Hall & Smith, 1991). However, this technique required relatively large amounts of non-degraded DNA and even so, no single or composite set of restriction enzymes have proved to be diagnostic for all subspecies (Meixner *et al.*, 2013). Later, restriction enzymes were employed in specific regions of the mtDNA such as CYTB (Crozier *et al.*, 1991), COX1 (Bouga *et al.*, 2005; Hall & Smith, 1991; Nielsen *et al.*, 2000; Stevanovic *et al.*, 2010), l-RNA (Hall & Smith, 1991; Ozdil & Ilhan, 2012a) and ND5 (Bouga *et al.*, 2005; Ozdil & Ilhan, 2012b) using the technique PCR-RFLP. Among the different PCR-RFLP assays, the most popular is the commonly known as *Dra*I test, developed by Garnery *et al.* (1993). The *DraI* test consists of PCR amplification of the highly polymorphic non-coding region located between the $tRNA^{leu}$ and cox2 genes (originally named COI-COII intergenic region), followed by digestion with *DraI*. The polymorphism of the $tRNA^{leu}$-COX2 region results from a combination of length variation with restriction site polymorphisms. This intergenic region is composed by two distinct nucleotide sequences, named P and Q elements. The P element has several forms (P_0, P_1, P_2) and the Q element can be repeated one to five times (Rortais *et al.*, 2011). Garnery *et al.* (1993) showed that the polymorphism existent in this region was able to allocate honey bees in one of the three lineages (A, African; M, Western European; C, Eastern European) previously described by Ruttner (1988) using morphometric data. Later, the Y lineage was proposed and the complexity within the lineages, particularly the African, increased. Franck *et al.* (2001) examined 738 colonies from Africa and classified the African diversity into three sub-lineages (A_I, A_{II} and A_{III}). Afterwards, Alburaki *et al.* (2011) proposed a new African sub-lineage, the Z.

Discrimination between A, M, and C lineages and African sub-lineages (A_I, A_{II}, A_{III}, and Z) has been based on *Dra*I sites located at the 3' end of $tRNA^{leu}$ and the 5' end of the first Q element, as well as on indels located in the P element. The C lineage is characterized by the absence of the P element whereas the M lineage contains a P element containing 54–56 bp. The African haplotypes carry the P_0 element with 67 bp (sub-lineages A_I, A_{II} and Z), and also the P_1 element (sub-lineage A_{III}), characterized by a 15-bp deletion at the 3' end of the P element. Sub-lineage A_{II} is differentiated from sub-lineage A_I by the absence of the *Dra*I site, at the 5' end of the first Q element and sub-lineage Z presents an additional *Dra*I site in the middle of the first Q element.

In addition to being very informative for lineage and sub-lineage discrimination, the tRNA^{leu}-cox2 intergenic region has other advantages: (1) there is a large catalogue of haplotypes developed from honey bees collected across the world (Collet *et al.*, 2006; Meixner *et al.*, 2013; Pinto *et al.*, 2012; Shaibi *et al.*, 2009); (2) haplotype identification by means of the *Dra*I test requires a small-sized laboratory equipped with basic equipment and it is cheap and fast. On the other hand, the data generated from this region is not the most appropriate for phylogenetic and phylogeographical inferences due the large number of indels and it is unclear whether this segment of mtDNA reflects the "real" maternal history. While all mitochondrial genes are linked, it has been shown that different regions evolve at different rates and sometimes lead to incongruent results (Duchene *et al.*, 2012; Keis *et al.*, 2013; Meiklejohn *et al.*, 2014; Sasaki *et al.*, 2005; Zardoya & Meyer, 1996). In honey bees, Ilyasov, Poskryakov, and Nikolenko (2016) showed that while ND2, ND4, ND4L, ND5, ND6, COX1 and COX3 allow subspecies differentiation, COX2, ATP6, ATP8, ND1, ND3 distort the phylogeny.

The honey bees from the Iberian Peninsula (*Apis mellifera iberiensis)* have been intensively surveyed with several molecular markers, including mtDNA, more specifically the *Dra*I test (Cánovas *et al.*, 2008; Franck *et al.*, 1998; Miguel *et al.*, 2007; Pinto *et al.*, 2013). These studies revealed the co-existence of the African (A) and Western European (M) lineages, forming a northeastern-southwestern cline. Moreover, the presence of the sub-lineages A_I, A_{II}, A_{III} has been detected within the African lineage. While sub-lineage A_{II} is the least frequent and is mainly found in the center of Spain, sub-lineage A_I is the most widespread, being more frequent in the south of Spain, and sub-lineage A_{III} occurred in high proportions in the north of Portugal.

Chávez-Galarza *et al.* (2017) carried out a comprehensive survey of maternal variation of the tRNA^{leu}-cox2, sequencing this region in 711 individuals collected across the entire Iberian honey bee range. From these 711, 87 individuals representing the diversity found in Iberian Peninsula were selected to be whole-genome sequenced (WGS). To this set of 87 Iberian specimens, we added 36 individuals from six other subspecies, including *A. m. mellifera*, *A. m. sahariensis*, *A. m. intermissa*, *A. m. carnica*, *A. m. ligustica* and *A. m. siciliana*. Using this comprehensive mtDNA-WGS dataset, we aimed to investigate if the mitogenome supports the information obtained by the tRNA^{leu}-cox2 intergenic region, namely: a) if individual genes provide the same topology as the mitogenome; b) if the three lineages occurring in Europe are supported, c) if the three African sub-lineages reported for Iberia are confirmed and, globally, d) if the diversity of each individual gene

and the mitogenome is concordant with the variation of the $tRNA^{leu}$-cox2 intergenic region found in Iberia.

Methods

Sampling

A total of 123 individuals representing seven subspecies, three lineages (A, M and C) and three African sub-lineages (A_I, A_{II} and A_{III}), were sampled. The *A. m. iberiensis* sample comprises a total of 87 haploid drones collected in 2010 from 16 sampling sites distributed throughout three north-south transects in Iberia: one along the Atlantic coast (AT: 31 individuals), one along the center (CT: 33 individuals), and another along the Mediterranean coast (MT: 23 individuals; Figure VII-1). Each individual represents a single colony and apiary. The remaining 36 individuals represent six subspecies, including eight *A. m. mellifera* (Denmark, France, Netherlands, Norway and Scotland), seven *A. m. sahariensis* (Algeria and Morocco), 12 *A. m. intermissa* (Algeria and Morocco), three *A. m. carnica* (Croatia and Serbia), four *A. m. ligustica* (Italy), and two *A. m. siciliana* (Sicily; Figure VII-1. and Table Sup VII-1).

Drones were taken from inside the hives, placed in absolute ethanol, and then stored at -20°C until DNA extraction. Genomic DNA was extracted from the thorax of the 123 individuals using a phenol/chloroform isoamyl alcohol (25:24:1) protocol (Sambrook *et al.*, 1989).

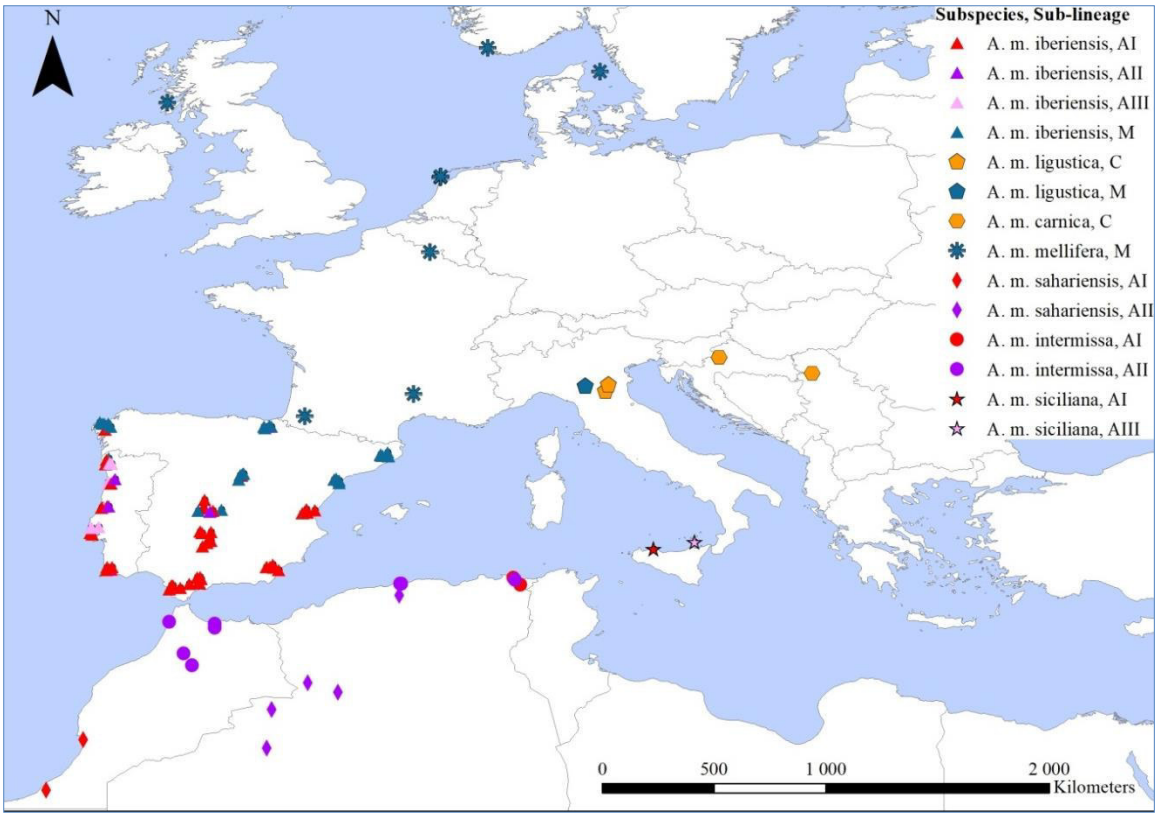


Figure VII-1 - Geographical location of 123 colonies of *A. m. iberiensis* (87 colonies), *A. m. ligustica* (4 colonies), *A. m. carnica* (3 colonies), *A. m. mellifera* (8 colonies), *A. m. sahariensis* (7 colonies), *A. m. intermissa* (12 colonies) and *A. m. siciliana* (2 colonies). Each point represents a colony and the color/symbol represents the subspecies and the lineages (A, M and C) or African sub-lineages (A_I, A_{II} and A_{III}) based on the $tRNA^{leu}$-cox2 region.

$tRNA^{leu}$-cox2 intergenic region

The haplotypes for 114 individuals (excluding *A. m. sahariensis* and *A. m. siciliana*) were obtained from the $tRNA^{leu}$-cox2 intergenic region sequenced by Chávez-Galarza *et al.* (2017). For the remaining nine individuals the $tRNA^{leu}$-cox2 intergenic region was PCR-amplified with the E2 and H2 primers with a slightly modified reaction and conditions established by Garnery *et al.* (1993). PCR products were directly sequenced in both directions using the Sanger method. The alignment of the sequences was performed with MEGA 6.06 (Tamura *et al.*, 2013) and the haplotypes were named following the nomenclature system revised by Chávez-Galarza *et al.* (2017). GENEALEX 6.5 (Peakall & Smouse, 2006) was used to estimate the mean number of haplotypes per locus (Na), effective number of haplotypes (Ne), number of private haplotypes (Np), and unbiased diversity (uh).

Mitogenome sequencing and filtering

Whole genome sequencing (WGS) was accomplished using the Illumina HiSeq 2500 platform. Sequence reads were mapped against the reference honey bee genome Amel_4.5 (see Henriques *et al.* 2018, for further details). Here, we only used the reads mapped against the mitochondrial DNA. After the SNP calling several filters were applied to reduce poor mapping and spurious positions. The SNPs that had more than two alleles, showed a quality score <50 and were present in <37 samples were removed from the dataset. MtDNA data was intentionally misspecified to be diploid in the SNP calling process. The heterozygous positions were removed from the dataset as they do not represent true SNPs. Some problematic sites were identified by more than one filter. Therefore, only 645 out of 795 identified SNPs were kept for further analysis (Table Sup VII-2). We used bcftools implemented in the SAMtools software package (Li *et al.*, 2009) to incorporate each of the 645 variants into the reference mitogenome.

To understand if different parts of the mitogenome provide similar results to the $tRNA^{leu}$-cox2 intergenic intergenic region, the mitogenome and each protein-coding and rRNAs genes were analyzed and compared against this region. The mitogenome consisted of nearly the complete mtDNA; the fragment missing is positioned within the $tRNA^{leu}$-cox2 intergenic region, corresponding to the P-element indel and the Q-element indels that characterize the reference *A. m. ligustica* (C lineage) mitogenome (Garnery *et al.*, 1993).

Population and phylogenetic analyses

Several summary statistics including nucleotide diversity (π), haplotype diversity (H), the number of haplotypes (Hd) and the number of polymorphic sites were calculated for each of the 16 datasets (mitogenome, individual protein-coding and rRNAs genes) using DnaSP 5.10 (Rozas *et al.*, 2003). Other molecular indexes, such as number of transitions, transversions and private substitution sites were calculated using Arlequin 3.5.1.2 (Excoffier *et al.*, 2005). The individuals were grouped by subspecies, and because some subspecies carry haplotypes belonging to different lineages, the individuals were further grouped by lineages. The average number of pairwise differences between populations (πXY) and within populations (πX) were calculated by Arlequin. To avoid spurious conclusions, due to the bias introduced when small sample sizes are small, diversity estimates for populations represented by <5 individual were not compared.

Phylogenetic relationships between haplotypes obtained from the mitogenome and from all individual protein coding and rRNAs genes were inferred using median-joining network analysis (Bandelt *et al.*, 1999) in PopART (http://popart.otago.ac.nz). Bayesian phylogenetic analysis was performed on mitogenome and each gene using MrBayes 3.2.6 (Ronquist & Huelsenbeck, 2003). The best-fitting evolutionary model for each gene and for the mitogenome was calculated according to the Bayesian information criterion (BIC) using the PartitionFinder v1.1.1 (Lanfear *et al.*, 2012). The selected models were run on MrBayes 3.2.6 in the CIPRES platform (Miller *et al.*, 2010). Four Bayesian independent runs with random starting trees were performed and posterior distributions of parameters, including the tree, were estimated using Markov chain Monte Carlo (MCMC) sampling. Samples were drawn every 100 MCMC steps over a total of 50,000,000 steps, with the first 125,000 trees discarded as burn-in. The homologous mitochondrial sequence of *Apis cerana* (GI:299829158), the sister species of *A. mellifera*, was used as an outgroup.

Dataset comparisons

For the mitogenome and each gene, two different methods were used to delimitate the different groups. In one method, the aligned sequences were analysed by Automatic Barcode Gap Discovery (ABGD), a program available at http://wwwabi.snv.jussieu.fr/public/abgd/abgdweb.html, with the default settings using either the Kimura-2-Parameter (K2P; Kimura, 1980) and Jukes-Cantor (JC; Jukes *et al.*, 1969) distance metrics. In the other method, the aligned sequences were analysed by Bayesian Poisson Tree Processes (bPTP; Zhang *et al.*, 2013), which delimits groups based on the phylogenetic species concept. We used the bayesian non-ultrametric phylograms as input obtained by MrBayes, which was submitted to the Exelixis Lab web-server (http://species.h-its.org/ptp/). The bPTP analysis was run for 500,000 MCMC generations, with a thinning value of 100 and a burn-in of 25%.

The phylogenetic topologies generated by MrBayes 3.2.6 from the mitogenome and each gene were compared using the APE package in R (R Development Core Team, 2011; Paradis *et al.*, 2004) and using the PH85 distance (Penny & Hendy, 1985). The pairwise tree distances were then used by Past 3.08 (Hammer *et al.*, 2001) to create a Neighbor-Joining (NJ) dendrogram, which represents gene groupings by topology similarities.

Results

Distribution of SNPs in the mitogenome

The sequencing depth of coverage of each individual ranged from 2,523 to 7,758X (Table Sup VII-1). A total of 795 SNPs were identified in the 16,343 bp reference mitochondrial genome; however, 150 SNPs did not pass one or more filtering criteria (Table Sup VII-2). The 645 SNPs were not uniformly distributed along the genome. The proportion ranges from 2.41% in l-rDNA until 5.66% in ATP8. Of the 645 SNPs, 506 were located in the coding region, being 421 transitions and 85 transversions and 151 were in non-synonymous positions. Gene ND4 contained the highest number of non-synonymous SNPs (24) whereas the shortest genes ATP8 (159 bp) and ND4L (264 bp) contained the lowest number (3 SNPs). ATP6 showed the highest proportion of non-synonymous SNPs (52%) and the gene COX1 contains the lowest proportion of non-synonymous SNPs (14 %; Table Sup VII-3).

tRNA^{leu}-cox2 intergenic region

Using the nomenclature system developed by Garnery *et al.* (1993) and revised by Rortais *et al*. for the M lineage and Chávez-Galarza *et al.* (2017) for the A lineage, a total of 30 haplotypes and 83 variants were identified in the 123 individulas. The most common haplotypes were A1 (14 individuals), A2 (19 individuals) and M4 (16 individuals) with six, eight and five variants each, respectively. The variation was grouped into lineages A (80 individuals), M (38 individuals) and C (five individuals). The 80 A-lineage individuals were further divided in sub-lineage A_I (50 individuals), A_{II} (20 individuals) and A_{III} (ten individuals; Table Sup VII-1).

The subspecies *A. m. sahariensis*, *A. m. intermissa*, *A. m. carnica*, *A. m. mellifera* and *A. m. siciliana* carried haplotypes from a single lineage whereas *A. m. iberiensis* and *A. m. ligustica* carried haplotypes belonging to two different lineages (Table VII-1).

The most frequent sub-lineage in *A. m. sahariensis* and *A. m. intermissa* was A_{II}; however, there were also four haplotypes belonging to A_I (two for each subspecies). Haplotypes of African ancestry carried by *A. m. iberiensis* belonged to sub-lineages A_I, A_{II} and $A_{III.}$ The most common and widespread haplotypes belonged to sub-lineage A_I (45 individuals) whereas haplotypes of sub-lineage A_{II} ancestry were the least common (five individuals). The haplotypes of sub-lineage A_{III} were mostly located in the Atlantic part of Iberia (eight out of nine; Figure VII-1).

Table VII-1 - Diversity measures for each subspecies within each lineage considering the variants (on the left) and the haplotypes (on the right) of the tRNA^{leu}-cox2 intergenic region.

Lineage	Populations	N[a]	#Na[b]	Private alleles	Ne[c]	Uh[d]
M	*A. m. iberiensis*	28	16 | 8	15 | 6	6.32 | 3.04	0.87 | 0.70
	A. m. mellifera	8	6 | 3	5 | 2	5.33 | 1.68	0.93 | 0.46
	A. m. ligustica	2	2 | 2	0 | 0	2.00 | 2.00	1.00 | 1.00
A	*A. m. iberiensis*	59	40 | 18	38 | 13	21.10 | 7.09	0.97 | 0.87
	A. m. intermissa	12	10 | 4	9 | 0	9.00 | 3.13	0.97 | 0.74
	A. m. sahariensis	7	7 | 3	7 | 0	7.00 | 2.33	1.00 | 0.67
	A. m. siciliana	2	2 | 2	2 | 1	2.00 | 2.00	1.00 | 1.00
C	*A. m. ligustica*	2	2 | 2	1 | 0	2.00 | 2.00	1.00 | 1.00
	A. m. carnica	3	3 | 2	1 | 0	3.00 | 1.80	1.00 | 0.67

[a] Number of individuals; [b] Mean number of haplotypes; [c] Number of effective haplotypes; [d] Unbiased haplotype diversity.

The haplotype diversity (Uh) using the variants inferred from the tRNA^{leu}-cox2 intergenic region indicates that *A. m. sahariensis* and *A. m. intermissa* (Uh=1 and 0.97, respectively) hold the highest diversity, whereas the lowest haplotype diversity was found for *A. m. iberiensis* of M-lineage ancestry (Uh=0.87). On the other hand, different diversity patterns were observed when only the haplotypes instead of the variants inferred from the tRNA^{leu}-cox2 intergenic region were assessed. Use haplotypes instead the variants resemble the results of the traditional *Dra*I test, where the haplotypes are inferred from a gel, so small differences are not detectable The A-lineage of *A. m. iberiensis* displayed the highest diversity (Uh=0.87), followed by *A. m. intermissa* (Uh=0.74) and the population with lowest values was *A. m. mellifera* (uh=0.46; Table VII-1).

Diversity across mitochondrial genes

A total of 115 haplotypes was inferred from the mitogenome (Table Sup VII-3). The genes with the highest number of haplotypes were CYTB (44), ND4 (43) and COX1 (41), and these were also the longest genes (>11,000 bp; Table Sup VII-3). When the length of the gene was accounted for, ATP8 (5.66%) and ND3 (5.65%) had the highest proportion of SNPs whereas rRNA genes had the lowest (2.41% and 3.18%; Table Sup VII-3). The number of haplotypes, the number of

parsimony-informative sites and the average number of nucleotide differences (k) were highly correlated with the length of the gene (r^2= 0.94, 0.99 and 0.99, respectively). The gene with the highest haplotype diversity was CYTB (1152 bp long; Hd=0.910) and the gene with the lowest is ATP8 (159 bp long; Hd=0.522). The gene with highest nucleotide diversity was ND6 (π=0.00935) and the lowest s-rDNA (0.0023; Table Sup VII-3).

Diversity across populations

The analysis of the complete mitogenome showed that in *A. m. iberiensis*, the M-lineage had higher diversity (π=22.707) than the A-lineage (π=10.854; Table VII-2). Indeed, *A. m. iberiensis* M-lineage had higher diversity than the other M-lineage subspecies *A. m. mellifera*. On the other hand, *A. m. iberiensis* belonging to A-lineage had lower diversity than *A. m. sahariensis* and *A. m. intermissa*. In the northern African populations, *A. m. sahariensis* had higher mitogenomic diversity (π=23.048) than *A. m. intermissa* (π=16.636), congruent with ND2, COX1 and ND3. Genes CYTB, ND1 and l-rRNA exhibited a diversity pattern similar to that of the variants from $tRNA^{leu}$-cox2 intergenic region, with *A. m. iberiensis* showing a lower diversity than the other populations of the corresponding lineage (Table Sup VII-4).

Table VII-2 - Mitogenome diversity measures. The individuals were divided by subspecies and for *A. m. iberiensis* and *A. m. ligustica* also by lineage.

Lineage	Populations	# Alleles	#TS[a]	#Tv[b]	Private alleles	π[c]
M	*A. m. iberiensis*	28	121	23	99	22.77
	A. m. mellifera	8	56	9	34	20.32
	A. m. ligustica	2	0	0	0	0.00
A	*A. m. iberiensis*	59	115	30	104	10.85
	A. m. intermissa	12	42	7	27	13.64
	A. m. sahariensis	7	67	5	42	23.05
	A. m. siciliana	2	31	0	16	31.00
C	*A. m. ligustica*	2	11	2	9	13.00
	A. m. carnica	3	22	7	23	19.33

[a] Number of transitions; [b] Number of transversions; [c] Nucleotide diversity.

Structure

The diversity patterns revealed by the mitogenome and the single genes were not totally concordant either with each other or with the $tRNA^{leu}$-cox2 intergenic region. Results from population average pairwise differences analysis show that all genes can differentiate the three lineages. Differences among genes arise in the relationship among the three lineages. Lineages M and C were the most divergent, as revealed by the majority of the genes and the mitogenome. In contrast, COX2 and CYTB show that lineages A and C were the most divergent and finally, for genes ND2 and ND6, the most divergent lineages were A and M (Table Sup VII-5).

The mitogenome and 9 out of 15 genes are able to distinguish not only the three lineages but also the different subspecies that share the same lineage. This is not the case for the genes with lower number of SNPs (ATP8 and ND4L) and genes ATP6, COX3, ND3 and s-rRNA where all subspecies of African ancestry had π=0.00 (Tables Sup VII-4 and Sup VII-5).

Phylogeographical structure

The network inferred from the mitogenome shows the distribution of 115 haplotypes in three main haplogroups, which corresponds to the three lineages supported by the *DraI* test (Figure VII-2). The African haplogroup can be further subdivided into three different clusters: one containing mostly North African individuals, one containing the Iberian and African individuals near to Strait of

Gibraltar, and the other containing individuals mostly located in Portugal (Figure VII-3). The M-lineage haplogroup can be also subdivided in four different clusters: two of them mostly confined to the Iberian Peninsula; one shared between the Iberian Peninsula and France; and one that is characteristic from the Northern European populations (Figure VII-3).

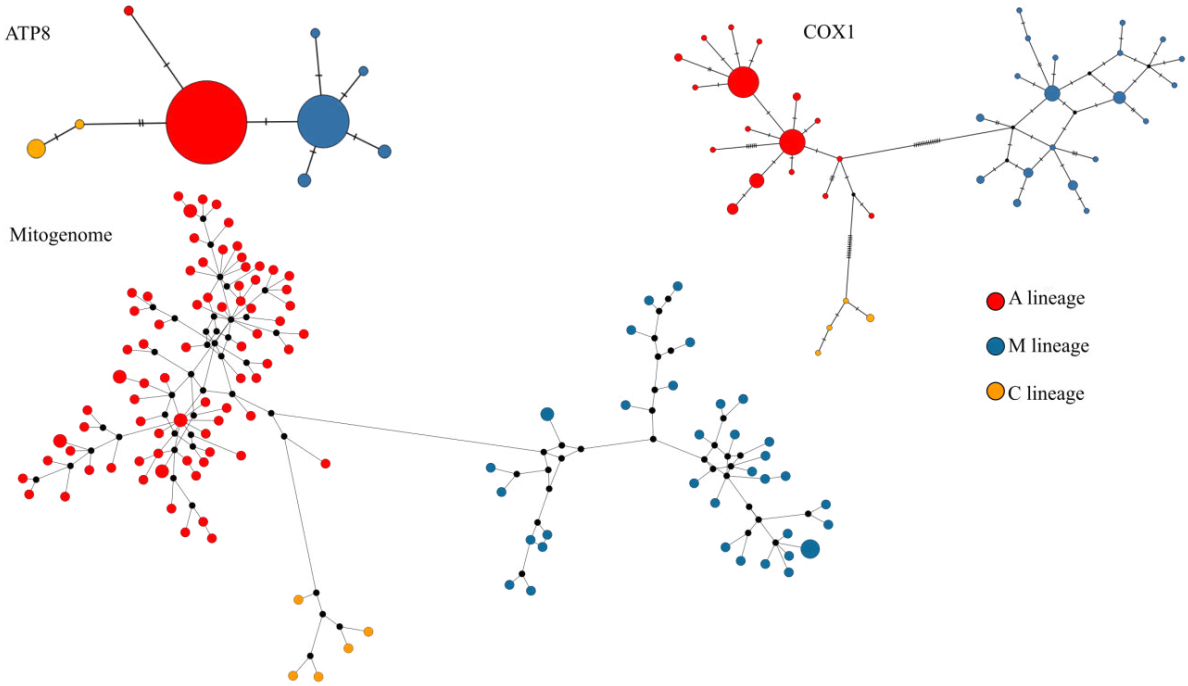


Figure VII-2 - Median-joining network using genes with different length and number of SNPs and the mitogenome. The ATP8 is the gene with lowest number of SNPs; COX1 is one of the most informative protein coding genes. Unsampled or extinct haplotypes are indicates as black circles. The size of circles is proportional to haplotype frequencies. Links between haplotypes are proportional to genetic distances between them. The colors correspond to the three lineages, as identified by the *Dra*I test.

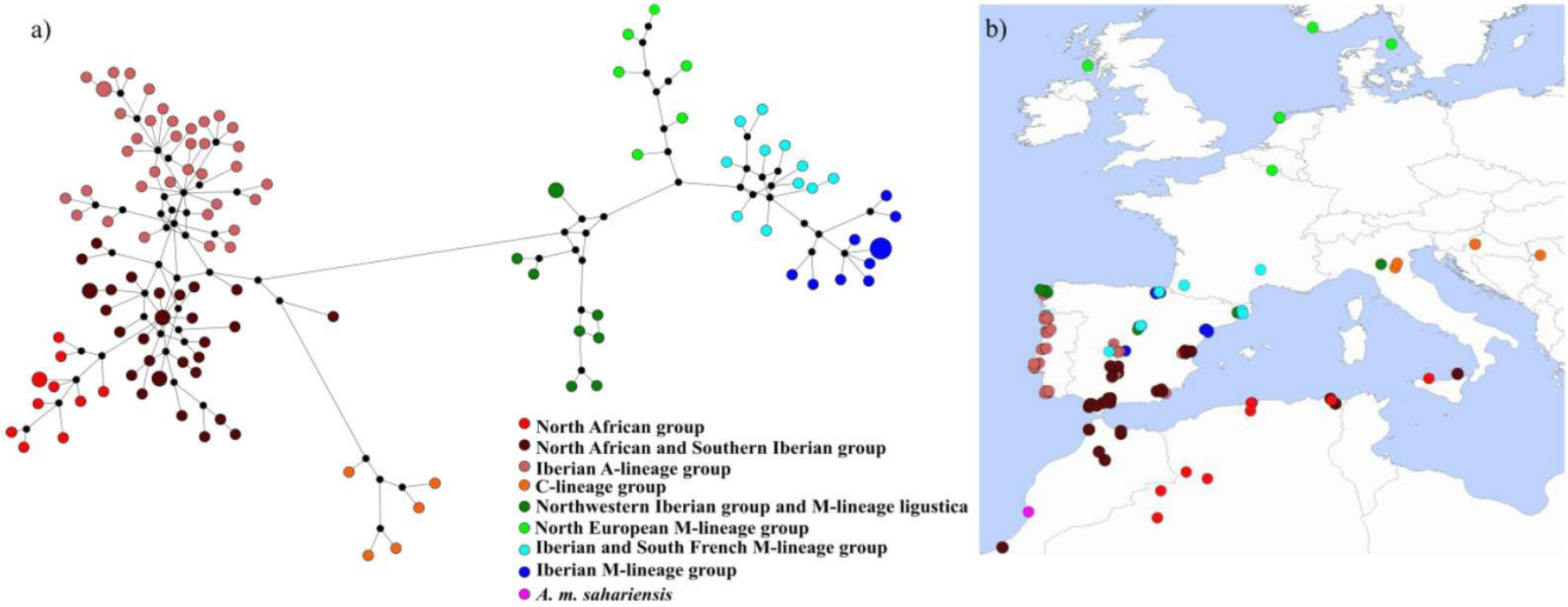


Figure VII - 3 - Phylogeographical relationship between the 123 samples. a) Median-joining network inferred from the mitogenome and colored by haplogroups. b) Geographical distribution of the haplogroups identified by the mitogenome network.

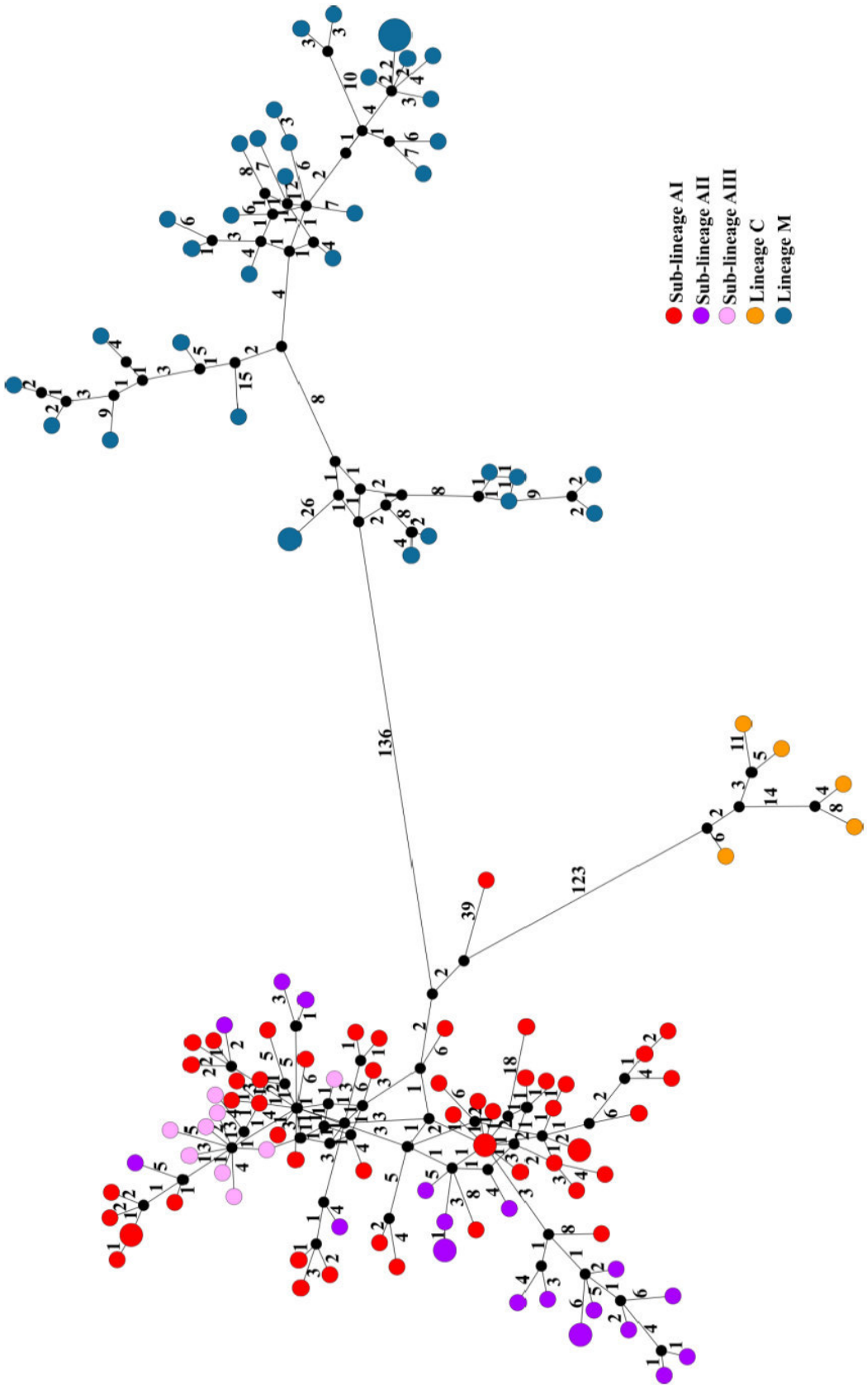


Figure VII - 4 - Median-joining network inferred from the mitogenome. The size of circles is proportional to haplotype frequencies. Links between haplotypes are proportional to genetic distances between them. The colors correspond to the lineages or African sub-lineages, as identified by the *Dra*I test.

While the haplogroups inferred from the mitogenome overlapped the lineages identified by the tRNA^{leu}-cox2 intergenic region, the subgroups do not support the African sub-lineages (Figure VII-4). Haplotypes of sub-lineages A_I and A_{II} ancestry were widespread across African clusters. Haplotypes of sub-lineage A_{III} ancestry were more confined in the network, although they were connected with other haplotypes belonging to sub-lineages A_I and A_{II}. The tRNA^{leu}-cox2 intergenic region showed that in the population from Africa and Iberia few individuals share haplotypes of the same sub-lineage. On the other hand, the mitogenome showed that the African individuals located nearby Strait of Gibraltar were related with the individuals from the southern Spain. Another difference was that while the tRNA^{leu}-cox2 intergenic region placed the populations from the north of Portugal in a separate clusters, the mitogenome placed all populations from Portugal together in the same cluster (Figure VII-1 and Figure VII-3).

The mitogenome provided greater phylogeographical resolution than individual genes (Figure VII-2). All genes, with the exception of s-rRNA, originated networks that divide the haplotypes in three haplogroups matching evolutionary lineages. The network inferred from the s-rRNA gene places the A- and M- haplotypes mostly in two separate haplogroups, but some A-lineage haplotypes were in the M-lineage group and vice-versa.

The resolution within each haplogroup depends on the length of the gene and number of SNPs; the largest genes containing a higher number of SNPs had higher resolution (Figure VII-2). In addition to different resolutions, in CYTB, COX2 and ND2, concordantly with the pairwise differences analysis, C-lineage individuals were closer to the M-lineage whereas in the rest of the datasets the C-lineage was closer to the A-lineage (Figure Sup VII-1).

The bayesian phylogenetic tree inferred from the mitogenome was concordant with the network analysis. In addition to the main division in three evolutionary lineages, which was also supported by the methods used to delimitate the different groups (bPTP and ABGD), the phylogenetic clades support the subgroups found in the network analysis in the A and M lineages (Table Sup VII-7). The only exception was found on the North African and Southern Iberian group that did not formed a clade in the phylogenetic tree (Figure VII-6). In the same tree it is possible to observe that the individuals with the same haplotype were not always clustered together. Conversely, the trend in the phylogenetic tree leans towards the formation of a cluster with individuals from the same geographical area, such as the case of the seven A1 haplotypes, five of them from the southern part of Portugal (AT8), clustered together were more related with one A9

from the same location (2277) than with the other two A1 individuals from the center of Portugal (2069 and 2345).

The Neighbour-Joining dendrogram represents the distance between the individual genes and the mitogenome. The Long Branch between the individual coding genes and the mitogenome shows that the topology of the mitogenome is remarkably different from the individual coding genes (Figure VII-5). The most similar gene to the mitogenome was COX1, although it was still very distant (PH85=53; Table Sup VII-6). In addition, the individual genes produced different results, leading to incongruent topologies. Only 3 pairs of genes (ATP8:ND4L, ND3:ATP6, ND6:COX3) with similar length were grouped (Figure VII-5 and Table Sup VII-3).

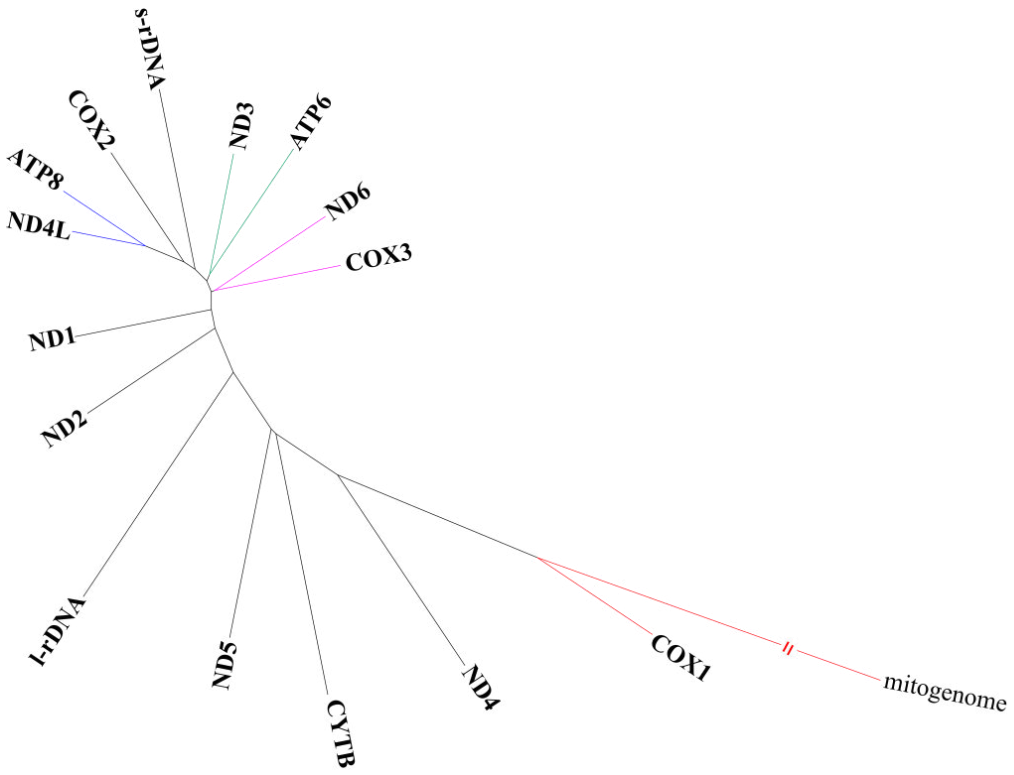


Figure VII-5 - Neighbour-joining dendrogram for the PH85 topology distance between the single genes and the mitogenome.

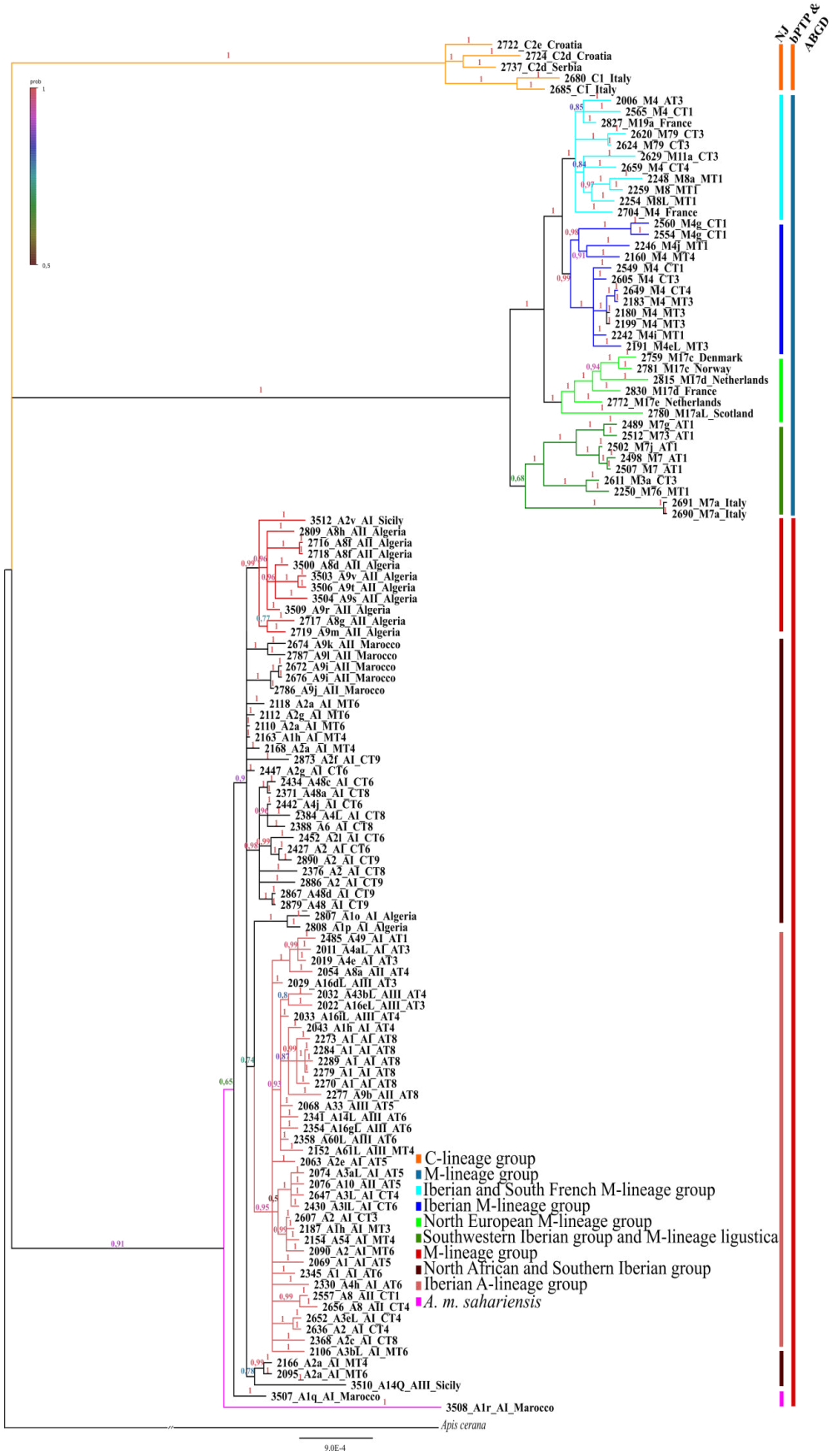


Figure VII-6 - Bayesian phylogenetic tree inferred from the mitogenome. Values indicate the bootstrap support and are colored from the less to the most probable. The first vertical bar depicts the partition concordance of the Bayesian and NJ analysis. The second vertical bar represents the results of the group delimitation analysis (bPTP and ABGD), which correspond to the evolutionary lineages C, M and A, represented by orange, blue and red.

The trees based on individual genes showed much lower resolution than the mitogenome. Most of the phylogenetic trees (with exception of AT8 and s-rRNA genes) split the haplotypes into the three main lineages; However, the resolution in each clade was very low. The bPTP and ABGD methods were not concordant in the number of groups for single genes (Table Sup VII-7). The bPTP was the most conservative method identifying for most of the genes a single group and for ND6 and ND4 three groups that correspond to the three evolutionary lineages. The ABGD method showed a variable number of clusters ranging from one in ND5 (using the Jukes Cantor distance metric) and ND4L (using both Jukes Cantor and Kimura distance metrics) to 36 in ND1. ND2 and COX1 were the only genes with a concordant number of groups with the mitogenome. The genes with a higher number of groups separate single individuals into different groups.

Discussion

Mitochondrial DNA has been widely applied to study the genetic diversity and population structure in honey bees, using a range of molecular methods and genes such as COX1 (Bouga *et al.*, 2011; Hall & Smith, 1991; Nielsen *et al.*, 2000; Stevanovic *et al.*, 2010), CYTB (Collins *et al.*, 2000; Crozier *et al.*, 1991), ND5 (Ozdil & Ilhan, 2012b), ND2 (Arias & Sheppard, 1996, 2005; Bouga *et al.*, 2005), and l-rRNA (Hall & Smith, 1991; Ozdil & Ilhan, 2012a). However, the most popular segment of the mtDNA has been the the $tRNA^{leu}$ and COX2 intergenic region with over 100 haplotypes described so far (Chávez-Galarza *et al.,* 2017; Meixner *et al.*, 2013; Rortais *et al.*, 2011).

While all mitochondrial genes are linked, it has been shown that different regions evolve at different rates and sometimes lead to incongruent results (Duchene *et al.*, 2012; Keis *et al.*, 2013; Meiklejohn *et al.*, 2014; Sasaki *et al.*, 2005; Zardoya & Meyer, 1996). In honey-bees, Collet *et al.* (2007) and Ferreira *et al.* (2009) showed that the l-rRNA and CYTB PCR-RFLP patterns supported the evolutionary branches defined by the *Dra*I restriction pattern of the $tRNA^{leu}$-cox2 intergenic region. However, Ilyasov, Poskryakov, and Nikolenko (2016) showed that different genes originate different phylogenies. While ND2, ND4, ND4L, ND5, ND6, COX1 and COX3 allow the differentiation of the honey bee subspecies, COX2, ATP6, ATP8, ND1, ND3 distort the phylogeny, suggesting that different parts of the honey bee mitogenome provide different information. Using a mtDNA-WGS dataset developed from 123 individuals, we have the opportunity to study the mitogenomic

phylogenies and phylogeography through analysis of individual genes and the complete mitogenome and to compare this data with that obtained from the tRNA^{leu}-cox2 intergenic region.

The 123 mitogenomes were sequenced at extremely high coverage in every sample ranging from 2,523X to 7,758X. The high coverage strengths the reliability of the 645 SNPs found along the chromosome. A similar number of SNPs was found by Mikheyev *et al.* (2015). The proportion of variable sites is not uniform across the mitogenome ranging from 2.41% until 5.66%, with the ribosomal genes with lower proportion. This pattern is in agreement with the slower evolution rate that has been documented for s-rRNA, being typically useful to understand the genetic diversity of higher taxonomical levels (Arif & Khan, 2009). The difference of variable sites along the genome explains in part the phylogenetic and phylogeographic incongruence among the genes

All genes, with the exception of s-rRNA, generated networks that clearly divide the haplotypes in three haplogroups matching the lineages identified by the *Dra*I restriction pattern of the tRNA^{leu}-cox2 intergenic region. However, the African sub-lineages were not supported by any of the genes neither by the delimitation methods. The only gene that grouped the sub-lineage A_{III} in a different cluster was the s-rRNA, which has the slowest rate of evolution. Even at the haplotypic level, the individuals with the same haplotype were not always clustered together. Conversely, the trend in the phylogenetic tree leans towards the formation of a cluster with individuals from the same geographical area. The same clustering pattern was already described by Ilyasov, Poskryakov, Petukhov*, et al.* (2016) that stated that the classification system used to describe the haplotypes of the tRNA^{leu}-cox2 intergenic region does not reflect the relationships between the haplotypes, bringing together genetically distant haplotypes and separating genetically closer ones. A possible explanation is the hypervariability of the tRNA^{leu}-cox2 intergenic region that leads to homoplasy.

The mitogenome provided greater phylogeographical resolution than individual genes. For a total of 123 individuals we found 115 haplotypes. This level of variation was also found by Wragg *et al.* (2016). The gene with a phylogenetic topology closer to the mitogenome was COX1, which is the genetic marker recommended for metazoan DNA barcoding (Rodrigues *et al.*, 2017). The results obtained by each gene were not concordant with those of the tRNA^{leu}-cox2 intergenic region, but they were also incongruent between each other. One example of this incongruence among the different genes is the position of the lineage C in relation to the A and M lineages. Other important

observation, mainly for the A-lineage individuals, is that even when the full mitogenome is used we cannot differentiate honey bee subspecies. The individuals from North Africa near to Strait of Gibraltar and from Southern Spain are grouped together independently of the subspecies, proposing a recent shared evolutionary history.

In this study, we showed that the popular $tRNA^{leu}$-cox2 intergenic region does not fully captures the evolutionary history of the mitogenome. While the mitogenome supports the three evolutionary lineages defined by the *Dra*I restriction pattern of the $tRNA^{leu}$-cox2 intergenic region, the sub-lineages and even the haplotypes are not supported. The $tRNA^{leu}$-cox2 intergenic region plays an important role in describing the genetic diversity in honey bees mainly because of the large catalogue of haplotypes around the world. However, due to the hipervariability this region is not suited to study the evolutionary relationships between the individuals.

References

Achilli, A., Olivieri, A., Pellecchia, M., Uboldi, C., Colli, L., Al-Zahery, N., Accetturo, M., Pala, M., Kashani, B. H., Perego, U. A., Battaglia, V., Fornarino, S., Kalamati, J., Houshmand, M., Negrini, R., Semino, O., Richards, M., Macaulay, V., Ferretti, L., Bandelt, H.-J., Ajmone-Marsan, P., & Torroni, A. (2008). Mitochondrial genomes of extinct aurochs survive in domestic cattle. *Current Biology, 18*(4), R157-R158. doi: https://doi.org/10.1016/j.cub.2008.01.019

Alburaki, M., Moulin, S., Legout, H., Alburaki, A., & Garnery, L. (2011). Mitochondrial structure of Eastern honeybee populations from Syria, Lebanon and Iraq. *Apidologie, 42*(5), 628-641. doi: 10.1007/s13592-011-0062-4

Arias, M. C., & Sheppard, W. S. (1996). Molecular phylogenetics of honey bee subspecies (*Apis mellifera* L) inferred from mitochondrial DNA sequence. *Molecular Phylogenetics and Evolution, 5*(3), 557-566. doi: 10.1006/mpev.1996.0050

Arias, M. C., & Sheppard, W. S. (2005). Phylogenetic relationships of honey bees (Hymenoptera : Apinae : Apini) inferred from nuclear and mitochondrial DNA sequence data. *Molecular Phylogenetics and Evolution, 37*(1), 25-35. doi: 10.1016/j.ympev.2005.02.017

Arif, I., & Khan, H. (2009). Molecular markers for biodiversity analysis of wildlife animals: a brief review. *Animal Biodiversity and Conservation, 32*(1), 9-17.

Avise, J. C., & Ellis, D. (1986). Mitochondrial DNA and the evolutionary genetics of higher animals [and discussion]. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 312*(1154), 325-342.

Bandelt, H. J., Forster, P., & Rohl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution, 16*(1), 37-48.

Boore, J. L., Macey, J. R., & Medina, M. (2005). Sequencing and comparing whole mitochondrial genomes of animals. *Methods Enzymol, 395*, 311-348. doi: 10.1016/S0076-6879(05)95019-2

Bouga, M., Alaux, C., Bienkowska, M., Buchler, R., Carreck, N. L., Cauia, E., Chlebo, R., Dahle, B., Dall'Olio, R., De la Rua, P., Gregorc, A., Ivanova, E., Kence, A., Kence, M., Kezic, N., Kiprijanovska, H., Kozmus, P., Kryger, P., Le Conte, Y., Lodesani, M., Murilhas, A. M., Siceanu, A., Soland, G., Uzunov, A., & Wilde, J. (2011). A review of methods for discrimination of honey bee populations as applied to European beekeeping. *Journal of Apicultural Research, 50*(1), 51-84. doi: 10.3896/ibra.1.50.1.06

Bouga, M., Harizanis, P. C., Kilias, G., & Alahiotis, S. (2005). Genetic divergence and phylogenetic relationships of honey bee *Apis mellifera* (Hymenoptera: Apidae) populations from Greece and Cyprus using PCR – RFLP analysis of three mtDNA segments. *Apidologie, 36*(3), 335-344. doi: 10.1051/apido:2005021

Brito, P. H., & Edwards, S. V. (2009). Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica, 135*(3), 439-455. doi: 10.1007/s10709-008-9293-3

Cánovas, F., De la Rúa, P., Serrano, J., & Galián, J. (2008). Geographical patterns of mitochondrial DNA variation in *Apis mellifera iberiensis* (Hymenoptera: Apidae). *Journal of Zoological Systematics and Evolutionary Research, 46*(1), 24-30. doi: 10.1111/j.1439-0469.2007.00435.x

Chávez-Galarza, J., Garnery, L., Henriques, D., Neves, C. J., Loucif-Ayad, W., Jonhston, J. S., & Pinto, M. A. (2017). Mitochondrial DNA variation of *Apis mellifera iberiensis*: further insights from a large-scale study using sequence data of the tRNAleu-cox2 intergenic region. *Apidologie, 48*(4), 533-544. doi: 10.1007/s13592-017-0498-2

Chávez-Galarza, J., Henriques, D., Johnston, J. S., Carneiro, M., Rufino, J., Patton, J. C., & Pinto, M. A. (2015). Revisiting the Iberian honey bee (*Apis mellifera iberiensis*) contact zone: maternal and genome-wide nuclear variations provide support for secondary contact from historical refugia. *Molecular Ecology, 24*(12), 2973-2992. doi: 10.1111/mec.13223

Collet, T., Arias, M. C., & Del Lama, M. A. (2007). 16S mtDNA variation in *Apis mellifera* detected by PCR-RFLP. *Apidologie, 38*(1), 47-54. doi: 10.1051/apido:2006056

Collet, T., Ferreira, K. M., Arias, M. C., Soares, A. E. E., & Del Lama, M. A. (2006). Genetic structure of Africanized honeybee populations (*Apis mellifera L.*) from Brazil and Uruguay viewed through mitochondrial DNA COI–COII patterns. *Heredity (Edinb), 97*(5), 329-335. doi: 10.1038/sj.hdy.6800875

Collins, A. M., Sheppard, W. S., & Shimanuki, H. (2000). A scientific note on the identification of honey bee semen using a mitochondrial DNA marker. *Apidologie, 31*(5), 595-596 %@ 0044-8435.

Crozier, R. H., & Crozier, Y. C. (1993). The mitochondrial genome of the honeybee *Apis-mellifera*-Complete sequence and genome organization. *Genetics, 133*(1), 97-117.

Crozier, Y. C., Koulianos, S., & Crozier, R. H. (1991). An improved test for africanized honeybee mitochondrial-DNA. *Experientia, 47*(9), 968-969. doi: 10.1007/bf01929894

Duchene, S., Frey, A., Alfaro-Núñez, A., Dutton, P. H., Thomas P. Gilbert, M., & Morin, P. A. (2012). Marine turtle mitogenome phylogenetics and evolution. *Molecular Phylogenetics and Evolution, 65*(1), 241-250. doi: http://dx.doi.org/10.1016/j.ympev.2012.06.010

Excoffier, L., Estoup, A., & Cornuet, J. M. (2005). Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics, 169*(3), 1727-1738. doi: 10.1534/genetics.104.036236

Ferreira, K. M., Silva, O. L. E., Arias, M. C., & Del Lama, M. A. (2009). Cytochrome-b variation in *Apis mellifera* samples and its association with COI-COII patterns. *Genetica, 135*(2), 149-155. doi: 10.1007/s10709-008-9264-8

Feutry, P., Berry, O., Kyne, P. M., Pillans, R. D., Hillary, R. M., Grewe, P. M., Marthick, J. R., Johnson, G., Gunasekera, R. M., Bax, N. J., & Bravington, M. (2017). Inferring contemporary and historical genetic connectivity from juveniles. *Molecular Ecology, 26*(2), 444-456. doi: 10.1111/mec.13929

Feutry, P., Kyne, P. M., Pillans, R. D., Chen, X., Naylor, G. J. P., & Grewe, P. M. (2014). Mitogenomics of the Speartooth Shark challenges ten years of control region sequencing. *BMC Evol Biol, 14*(1). doi: 10.1186/s12862-014-0232-x

Filipi, K., Marková, S., Searle, J. B., & Kotlík, P. (2015). Mitogenomic phylogenetics of the bank vole *Clethrionomys glareolus*, a model system for studying end-glacial colonization of Europe. *Molecular Phylogenetics and Evolution, 82*(Part A), 245-257. doi: https://doi.org/10.1016/j.ympev.2014.10.016

Franck, P., Garnery, L., Loiseau, A., Oldroyd, B. P., Hepburn, H. R., Solignac, M., & Cornuet, J. M. (2001). Genetic diversity of the honeybee in Africa: microsatellite and mitochondrial data. *Heredity, 86*, 420-430. doi: 10.1046/j.1365-2540.2001.00842.x

Franck, P., Garnery, L., Solignac, M., & Cornuet, J.-M. (1998). The Origin of west european subspecies of honeybees (*Apis mellifera*): new insights from microsatellites and mitochondrial data. *Evolution, 52*(4), 1119-1134. doi: 10.1111/j.1558-5646.1998.tb01839.x

Garnery, L., Cornuet, J. M., & Solignac, M. (1992). Evolutionary history of the honey bee *Apis mellifera* inferred from mitochondrial DNA analysis. *Molecular Ecology, 1*(3), 145-154. doi: 10.1111/j.1365-294X.1992.tb00170.x

Garnery, L., Solignac, M., Celebrano, G., & Cornuet, J. M. (1993). A simple test using restricted PCR-amplified mitochondrial DNA to study the genetic structure of *Apis mellifera L*. *Experientia, 49*(11), 1016-1021. doi: 10.1007/bf02125651

Gilbert, M. T. P., Drautz, D. I., Lesk, A. M., Ho, S. Y. W., Qi, J., Ratan, A., Hsu, C.-H., Sher, A., Dalén, L., Götherström, A., Tomsho, L. P., Rendulic, S., Packard, M., Campos, P. F., Kuznetsova, T. V., Shidlovskiy, F., Tikhonov, A., Willerslev, E., Iacumin, P., Buigues, B., Ericson, P. G. P., Germonpré, M., Kosintsev, P., Nikolaev, V., Nowak-Kemp, M., Knight, J. R., Irzyk, G. P., Perbost, C. S., Fredrikson, K. M., Harkins, T. T., Sheridan, S., Miller, W., & Schuster, S. C. (2008). Intraspecific phylogenetic analysis of Siberian woolly mammoths using complete mitochondrial genomes. *Proceedings of the National Academy of Sciences, 105*(24), 8327-8332. doi: 10.1073/pnas.0802315105

Hall, H. G., & Smith, D. R. (1991). Distinguishing African and European honeybe matrilines using amplified mitochondrial DNA. *Proceedings of the National Academy of Sciences of the United States of America, 88*(10), 4548-4552. doi: 10.1073/pnas.88.10.4548

Hammer, Ø., Harper, D. A. T., & Ryan, P. D. (2001). Paleontological statistics software: package for education and data analysis. *Palaeontologia Electronica*(4).

Hong, Y., Duo, H., Hong, J., Yang, J., Liu, S., Yu, L., & Yi, T. (2017). Resequencing and comparison of whole mitochondrial genome to gain insight into the evolutionary status of the Shennongjia golden snub-nosed monkey (SNJ *R. roxellana*). *Ecology and Evolution, 7*(12), 4456-4464. doi: 10.1002/ece3.3011

Ilyasov, R. A., Poskryakov, A. V., & Nikolenko, A. G. (2016). Seven genes of mitochondrial genome enabling differentiation of honeybee subspecies *Apis mellifera*. *Russian Journal of Genetics, 52*(10), 1062-1070. doi: 10.1134/s1022795416090064

Ilyasov, R. A., Poskryakov, A. V., Petukhov, A. V., & Nikolenko, A. G. (2016). New approach to the mitotype classification in black honeybee *Apis mellifera mellifera* and Iberian honeybee *Apis mellifera iberiensis*. *Russian Journal of Genetics, 52*(3), 281-291. doi: 10.1134/s1022795416020058

Jacobsen, M. W., Hansen, M. M., Orlando, L., Bekkevold, D., Bernatchez, L., Willerslev, E., & Gilbert, M. T. P. (2012). Mitogenome sequencing reveals shallow evolutionary histories and recent divergence time between morphologically and ecologically distinct European whitefish (*Coregonus spp.*). *Molecular Ecology, 21*(11), 2727-2742. doi: 10.1111/j.1365-294X.2012.05561.x

Jukes, T. H., Cantor, C. R., & Munro, H. N. (1969). Evolution of protein molecules. *Mammalian protein metabolism, 3*(21), 132.

Keis, M., Remm, J., Ho, S. Y. W., Davison, J., Tammeleht, E., Tumanov, I. L., Saveljev, A. P., Männil, P., Kojola, I., Abramov, A. V., Margus, T., & Saarma, U. (2013). Complete mitochondrial genomes and

a novel spatial genetic method reveal cryptic phylogeographical structure and migration patterns among brown bears in north-western Eurasia. *Journal of Biogeography, 40*(5), 915-927. doi: 10.1111/jbi.12043

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution, 16*(2), 111-120 %@ 0022-2844.

Lanfear, R., Calcott, B., Ho, S. Y. W., & Guindon, S. (2012). PartitionFinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution, 29*(6), 1695-1701. doi: 10.1093/molbev/mss020

Li, H., Gao, J., Liu, H., & Cai, W. (2009). Progress in the researches on insect mitochondrial genome and analysis of gene order. *Science Foundation in China, 17*(2), 39-45. doi: 10.1088/1005-0841/17/2/004

Ma, C., Yang, P., Jiang, F., Chapuis, M.-P., Shali, Y., Sword, G. A., & Kang, L. E. (2012). Mitochondrial genomes reveal the global phylogeography and dispersal routes of the migratory locust. *Mol Ecol, 21*(17), 4344-4358. doi: 10.1111/j.1365-294X.2012.05684.x

Meiklejohn, K. A., Danielson, M. J., Faircloth, B. C., Glenn, T. C., Braun, E. L., & Kimball, R. T. (2014). Incongruence among different mitochondrial regions: a case study using complete mitogenomes. *Molecular Phylogenetics and Evolution, 78*, 314-323 1055-7903.

Meixner, M. D., Pinto, M. A., Bouga, M., Kryger, P., Ivanova, E., & Fuchs, S. (2013). Standard methods for characterising subspecies and ecotypes of *Apis mellifera*. *Journal of Apicultural Research, 52*(4). doi: 10.3896/ibra.1.52.4.05

Miguel, I., Iriondo, M., Garnery, L., Sheppard, W. S., & Estonba, A. (2007). Gene flow within the M evolutionary lineage of *Apis mellifera*: role of the Pyrenees, isolation by distance and post-glacial re-colonization routes in the western Europe. *Apidologie, 38*(2), 141-155. doi: 10.1051/apido:2007007

Mikheyev, A. S., Tin, M. M. Y., Arora, J., & Seeley, T. D. (2015). Museum samples reveal rapid evolution by wild honey bees exposed to a novel parasite. *Nature Communications, 6*. doi: 10.1038/ncomms8991

Miller, M. A., Pfeiffer, W., & Schwartz, T. (2010, 14-14 Nov. 2010). *Creating the CIPRES Science Gateway for inference of large phylogenetic trees.* Paper presented at the 2010 Gateway Computing Environments Workshop (GCE).

Morin, P. A., Archer, F. I., Foote, A. D., Vilstrup, J., Allen, E. E., Wade, P., Durban, J., Parsons, K., Pitman, R., Li, L., Bouffard, P., Abel Nielsen, S. C., Rasmussen, M., Willerslev, E., Gilbert, M. T. P., &

Harkins, T. (2010). Complete mitochondrial genome phylogeographic analysis of killer whales (*Orcinus orca*) indicates multiple species. *Genome Research, 20*(7), 908-916. doi: 10.1101/gr.102954.109

Nielsen, D. I., Ebert, P. R., Page, R. E., Hunt, G. J., & Guzmán-Novoa, E. (2000). Improved polymerase chain reaction-based mitochondrial genotype assay for identification of the africanized honey bee (Hymenoptera : Apidae). *Annals of the Entomological Society of America, 93*(1), 1-6. doi: 10.1603/0013-8746(2000)093[0001:ipcrbm]2.0.co;2

Ozdil, F., & Ilhan, F. (2012a). Diversity of *Apis mellifera* Subspecies from Turkey revealed by sequence analysis of mitochondrial 16s rDNA region. *Biochemical Genetics, 50*(9-10), 748-760. doi: 10.1007/s10528-012-9517-1

Ozdil, F., & Ilhan, F. (2012b). Genetic Divergence of Turkish *Apis mellifera* Subspecies Based on Sequencing of ND5 Mitochondrial Segment. *Sociobiology, 59*(1), 225-234.

Pang, J.-F., Kluetsch, C., Zou, X.-J., Zhang, A.-b., Luo, L.-Y., Angleby, H., Ardalan, A., Ekström, C., Sköllermo, A., Lundeberg, J., Matsumura, S., Leitner, T., Zhang, Y.-P., & Savolainen, P. (2009). mtDNA data indicate a single origin for dogs south of Yangtze river, less than 16,300 Years Ago, from numerous wolves. *Molecular Biology and Evolution, 26*(12), 2849-2864. doi: 10.1093/molbev/msp195

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics, 20*(2), 289-290. doi: 10.1093/bioinformatics/btg412

Peakall, R. O. D., & Smouse, P. E. (2006). GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Resources, 6*(1), 288-295 1471-8286.

Penny, D., & Hendy, M. D. (1985). The use of tree comparison metrics. *Systematic Zoology, 34*(1), 75-82. doi: 10.2307/2413347

Pinto, M. A., Henriques, D., Guedes, H., Munoz, I., Azevedo, J., & De la Rúa, P. (2013). Maternal diversity patterns of Ibero-Atlantic populations reveal further complexity of Iberian honeybees. *Apidologie 44* 430-439.

Pinto, M. A., Muñoz, I., Chávez-Galarza, J., & De la Rúa, P. (2012). The Atlantic side of the Iberian Peninsula: a hot-spot of novel African honey bee maternal diversity. *Apidologie, 43*(6), 663-673. doi: 10.1007/s13592-012-0141-1

Pinto, M. A., Rubink, W. L., Coulson, R. N., Patton, J. C., & Johnston, J. S. (2004). Temporal pattern of Africanization in a feral honeybee population from Texas inferred from mitochondrial DNA. *Evolution, 58*(5), 1047-1055.

Rodrigues, M. S., Morelli, K. A., & Jansen, A. M. (2017). Cytochrome c oxidase subunit 1 gene as a DNA barcode for discriminating *Trypanosoma cruzi* DTUs and closely related species. *Parasites & Vectors, 10*, 488. doi: 10.1186/s13071-017-2457-1

Ronquist, F., & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics, 19*(12), 1572-1574.

Rortais, A., Arnold, G., Alburaki, M., Legout, H., & Garnery, L. (2011). Review of the DraI COI-COII test for the conservation of the black honeybee (*Apis mellifera mellifera*). *Conservation Genetics Resources, 3*(2), 383-391. doi: 10.1007/s12686-010-9351-x

Rozas, J., Sanchez-DelBarrio, J. C., Messeguer, X., & Rozas, R. (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics, 19*(18), 2496-2497.

Ruttner, F. (1988). *Biogeography and taxonomy of honeybees*: Springer Science & Business Media.

Sambrook, J., Fritsch , E. F., & Maniatis, T. (1989). *Molecular Cloning, 2nd edn.*

Sasaki, T., Nikaido, M., Hamilton, H., Goto, M., Kato, H., Kanda, N., Pastene, L., Cao, Y., Fordyce, R., Hasegawa, M., & Okada, N. (2005). Mitochondrial phylogenetics and evolution of mysticete whales. *Syst Biol, 54*(1), 77-90. doi: 10.1080/10635150590905939

Shaibi, T., Muñoz, I., Dall'Olio, R., Lodesani, M., De la Rúa, P., & Moritz, R. F. A. (2009). *Apis mellifera* evolutionary lineages in Northern Africa: Libya, where orient meets occident. *Insectes Sociaux, 56*(3), 293-300. doi: 10.1007/s00040-009-0023-3

Stevanovic, J., Stanimirovic, Z., Radakovic, M., & Kovacevic, S. R. (2010). Biogeographic study of the honey bee *(Apis mellifera* L.) from Serbia, Bosnia and Herzegovina and Republic of Macedonia based on mitochondrial DNA analyses. *Russian Journal of Genetics, 46*(5), 603-609. doi: 10.1134/s1022795410050145

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol, 30*(12), 2725-2729. doi: 10.1093/molbev/mst197

Techer, M. A., Clemencet, J., Simiand, C., Preeaduth, S., Azali, H. A., Reynaud, B., & Delatte, H. (2017). Large-scale mitochondrial DNA analysis of native honey bee *Apis mellifera* populations reveals a new African subgroup private to the South West Indian Ocean islands. *Bmc Genetics, 18*. doi: 10.1186/s12863-017-0520-8

Wang, Z., Shen, X., Liu, B., Su, J., Yonezawa, T., Yu, Y., Guo, S., Ho, S. Y. W., Vilà, C., Hasegawa, M., & Liu, J. (2010). Phylogeographical analyses of domestic and wild yaks based on mitochondrial DNA: new data and reappraisal. *Journal of Biogeography, 37*(12), 2332-2344. doi: 10.1111/j.1365-2699.2010.02379.x

Winkelmann, I., Campos, P. F., Strugnell, J., Cherel, Y., Smith, P. J., Kubodera, T., Allcock, L., Kampmann, M.-L., Schroeder, H., Guerra, A., Norman, M., Finn, J., Ingrao, D., Clarke, M., & Gilbert, M. T. P. (2013). Mitochondrial genome diversity and population structure of the giant squid *Architeuthis*: genetics sheds new light on one of the most enigmatic marine species. *Proceedings of the Royal Society B: Biological Sciences, 280*(1759). doi: 10.1098/rspb.2013.0273

Wragg, D., Marti-Marimon, M., Basso, B., Bidanel, J.-P., Labarthe, E., Bouchez, O., Le Conte, Y., & Vignal, A. (2016). Whole-genome resequencing of honeybee drones to detect genomic selection in a population managed for royal jelly. *Sci Rep, 6*, 27168. doi: 10.1038/srep27168

Zardoya, R., & Meyer, A. (1996). Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates. *Mol Biol Evol, 13*(7), 933-942.

Zhang, J., Kapli, P., Pavlidis, P., & Stamatakis, A. (2013). A general species delimitation method with applications to phylogenetic placements. *Bioinformatics, 29*(22), 2869-2876 1460-2059.

Chapter VIII.

Final discussion and concluding remarks

Final discussion and concluding remarks

In Europe there are 10 endemic subspecies that belong to the M and C evolutionary lineages. The European M-lineage includes only two subspecies: the Dark honey bee (*A. m. mellifera*) and the Iberian honey bee (*A. m. iberiensis*), the target subspecies of this work. The main goals of this dissertation were reveal the genetic structure of one of the most complex and diverse subspecies in Europe, the Iberian honey bee (*A. m. iberiensis*), and to develop molecular tools for estimating C-lineage introgression in the Iberian honey bee and its sister subspecies, the Dark honey bee (*A. m. mellifera*).

A. m. mellifera has an extensive native distribution. However its genetic integrity is severely compromised by introgressive hybridization in large areas throughout Northern and Western Europe. Across Different conservation programs and protected areas have been implemented in European in an attempt to protect and bring back the endangered Dark honey bee *A. m. mellifera*. The goal of chapter III was to develop a molecular tool capable of supporting breeding and management decisions in conservation areas. Therefore, five nested panels with 48, 96, 144, 192 and 384 ancestry informative SNPs were designed (fitted the plexes of GoldenGate® Assays formerly genotyped with the VeraCode® technology) and optimized to estimate introgression of C-lineage (*A. m. ligustica* and *A. m. carnica*) into the M-lineage (*A. m. mellifera*). As a baseline 1183 SNPs were used to genotype 113 haploid honey bees from eight countries using Illumina's BeadArray Technology and the Illumina GoldenGate® Assay with a custom Oligo Pool Assay (Illumina, San Diego, CA, USA). The ancestry-informative SNPs were simultaneously selected from the 1183 SNPs by pairwise Weir & Cockerham's F_{ST}, F_{ST}-based outlier test, Delta, I_n and PCA, to balance out the limitations of each individual method. Despite the different number of SNPs in each panel, all of them produced estimates of C-lineage proportions into *A. m. mellifera* greatly concordant with those inferred from the initial 1183 SNP dataset, as indicated by the high correlation values ($r \geq 0.997$).

Different motivations drove the chapter IV, a follow up of the chapter III: The performance of the panels developed in chapter III were assessed against WG resequencing data (2.399 million SNPs) and using a set of individuals, including F1 hybrids, obtained from controlled crosses purposely performed for this study. The Illumina's GoldenGate® genotyping technology was discontinued so the 144-SNP panel constructed in chapter III was used as baseline to design four

multiplexed assays to be genotyped using the iPLEX MassARRAY system of Agena BioScience™, which is amongst the most cost-effective genotyping platform for medium SNP throughput and high sample throughput. Another important motivation was the need of a standard method to describe the purity of honey bee populations in a wide geographical area to understand which conservation strategy is more efficient, and to target regions of greatest concern and greatest possible reward. Finally, the sensitivity of the system to be applied in pools of tissue or DNA was tested. This is particularly relevant for a species like the honey bee with its highly polyandrous mating system. Honey bee queens mate in flight with up to 20 drones. This means that in areas where *A. m. mellifera* and commercial colonies are sympatric, matings may occur with drones of C-lineage ancestry resulting in colonies made up of subfamilies with diverse genetic backgrounds some of which admixed. In cases of admixed gene pools, the analysis of a single honey bee may not represent the entire colony, therefore it is not adequate to support decision making. A ready-to-use accurate and cost-effective tool was provided with all genomic information along with the PCR and iPLEX primers for 117 high-quality SNPs multiplexed in the four assays for immediate application in genetic surveys and conservation management of *A. m. mellifera*. In addition, the dataset with the genotypes for haploid and diploid individuals of *A. m. mellifera*, *A. m. carnica* and *A. m. ligustica* were provided. These individuals can be used by others in introgression analysis as baseline reference populations with no need of inter-laboratory calibration. The final 117 SNPs are distributed into four panels M1=34, M2=32, M3=28, and M4=23 SNPs and across the 16 honey bee linkage groups. Only the assay M1 approached the maximum plexing capacity (40 SNPs), this drawback is in part due to the relatively small size of the baseline SNP set from which the Assay Design software had to work with. Nevertheless, the assays can be expanded to maximum capacity, using the *Replex* option of the Assay Design software. Additional nuclear SNPs, which can be chosen amongst top-ranked SNPs identified from WG by Parejo *et al.*, (2016), as well as mitochondrial SNPs can be added to the customized four assays for detecting C-derived genes at both genetic compartments.

In contrast with *A. m. mellifera*, *A. m. iberiensis* is still not threatened by introgression and the results obtained with sNMF and PCAdapt using 1,289,449 SNPs (Figure V-2) support the purity of *A. m. iberiensis* with a complex diversity pattern, showing that modern beekeeping has not disrupted its natural pattern of variation.

The complex diversity patterns found in *A. m. iberiensis* can be explained by neutral processes but also by selection. In chapter V the WG of the 87 Iberian honey bees were scanned for selection signals using three conceptually different methods (Samβada, LFMM and PCAdapt) and two datasets (genomic dataset and combination of a genomic and environmental dataset). Candidate genes that were simultaneously detected by at least two methods were further examined using the haplotype-based method iHS and protein modelling. A total of 830 SNPs distributed across 181 genes and 145 intergenic regions were cross-detected. There is a growing body of evidence that local adaptation is not primarily driven by protein-coding sequences, but rather that regulatory mutations might play a disproportionate larger part in the evolution of quantitative traits and of responses to environmental factors, such as stressors, resources and pathogens. The results were consistent with this idea, being the majority of the outlier SNPs (90.2%) beyond the exome. Moreover, using both genetic and environmental data genomic regions putatively under climate-driven adaptation were identified. These regions may allow the organism to adapt to different environmental conditions in an efficient way, and gene regulation is the most efficient and fastest way for an organism to adapt to different environments (López-Maury *et al.*, 2008).

Nevertheless, a significant enrichment of SNPs in coding regions was also observed, although only 40 SNPs were non-synonymous. However, there is a good chance that they are causal mutations, especially those in genes GB40077 and GB55263 (related with lipid biosynthesis), and GB45499 (membrane protein) as they could be linked to amino acid positions important for protein functioning.

The 181 candidate genes are involved in numerous different biological processes, some of them are membrane-related (one of the most enriched GO term) and others related with circadian clock genes (some of them cross-detected by all outlier tests). The first step for an organism to maintain homeostasis and to respond to the environment is to perceive the extracellular signals and translate them into cellular response, a role played by receptors and transport proteins of the membrane. The importance of membrane proteins for adaptation to new environments is evidenced by their rapid evolution compared with cytosolic proteins due to stronger adaptive selection to changing environments (Sojo *et al.*, 2016). Membrane proteins allow the interaction of cells with the environment mediating their behaviour and regulating their patterns of gene expression (Alberts, 2008; Hedin *et al.*, 2011; Reizer *et al.*, 1994). The circadian clock genes are also very important for an organism to anticipate and adapt to predictable environmental changes.

Because the environmental conditions are variable, clock systems have some plasticity (Yerushalmi & Green, 2009), however, with the rapid global changes characterized by unpredictability, the circadian and seasonal rhythms could be impaired and phenology changes are described as one of the first effects of global climate change (Helm *et al.*, 2013; Peñuelas & Filella, 2001).

Several lines of evidence link circadian rhythms with metabolism and feeding regimens (Froy & Miskin, 2007). The interdependence of the circadian clock and metabolism is very important in temperate zones where climate and food availability reflect seasonal environment variations and the organisms must adapt their behaviour to have a higher survival rate (Arrese & Soulages, 2010; Yurgel *et al.*, 2015). The honey bee relies on a circadian clock to synchronize foraging behaviour and reproductive swarming with the maximum daily and seasonal availability of food resources (Bloch, 2010; Simpson, 1958). During favourable seasons, honey bees collect honey for the next winter, as during winter they consume the harvested honey and use their metabolic heat to provide warmth to the colony. Following the winter pause, the Iberian honey bee activity starts earlier in the south than in the north of Iberia because suitable climate and food resources (pollen and nectar) are available earlier. Translocation experiments, involving several honey bees subspecies from Europe and *A. m. iberiensis* in Iberia have shown that local honey bees usually perform better and have longer survivorship than those introduced (Büchler *et al.*, 2014; Dražić *et al.*, 2014; Louveaux *et al.*, 1966).

The multicollinearity among the environmental variables makes difficult to identify causal selective pressures. However, it is interesting to notice that many of the candidate genes putatively under climate-driven adaptation allow the organism to sense and fine-tune with environmental oscillations.

While *A. m. iberiensis* is still not endangered, conservation measures should be applied before unique combinations of traits shaped by natural selection are irremediably lost, a possibility if Spanish and Portuguese beekeepers adopt a strategy of importing commercial C-lineage strains. Chapter VI was motivated with this in mind and, therefore, cost-effective and robust reduced SNP assays were constructed from 176 WGS to detect C-lineage introgression into *A. m. iberiensis*. Moreover, this panel can be applied to the Canary Islands, the archipelago of Azores and Madeira where there are signals of hybridization between C-lineage and *A. m. iberiensis*.

In addition to the construction of SNP assays to estimate introgression in *A. m. iberiensis,* the approach applied in chapter VI represents a rigorous methodological example that can be

applied for developing reduced SNP assays in any other organism. The chapter VI addressed also the importance of avoiding the long-standing problem of ascertainment bias: Taking advantage of the large and comprehensive WG dataset for the Iberian honey bee, the effect of sample size and sampling a geographically restricted area on detecting fixed SNPs was tested. The results showed that a bias on the number of fixed SNPs is introduced when sample size is small (N≤10) and sampling only captures a limited part of the genetic diversity.

The SNP assays developed in chapters III, IV and VI are a valuable tool to detect C-lineage introgression, a major concern in the fight to safeguard the reservoirs of unique combinations of genes and adaptations in *A. m. mellifera* and *A. m. iberiensis.* However, it should be noted, that these reduced assays are not suitable for standard population genetic analyses, including determining allelic diversity or measuring isolation by distance, genetic drift or bottleneck effect. The bias introduced through selection for markers that segregate among target populations would seriously compromise these calculations.

The maternal honey bee genetic variation has been widely accessed using the highly polymorphic $tRNA^{leu}$ -cox2 intergenic giving invaluable insights to characterize different populations. However, it is important to understand if this region is reliable for historical inference, which is the main goal of the chapter VII, where a 123-sample mtDNA-WGS dataset was used. The results showed that the popular $tRNA^{leu}$-cox2 intergenic region does not represent the evolutionary history of the mitogenome. While the mitogenome analysis supports the three evolutionary lineages defined by the *Dra*I restriction pattern of the $tRNA^{leu}$-cox2 intergenic region, it does not support the existence of different African sub-lineages. In addition, different parts of the genome provided distinct results, implying that the conclusions drawn from studies using only one locus need to be taken with caution. The best option to study the maternal phylogeographical history of an organism is to use the entire mitogenome since it exhibited far greater phylogeographical resolution than individual genes. Using the data of this chapter the next step is dissecting the Iberian honey bee complexities estimating genealogies, divergence events and possible adaptation to selective gradients.

The outcome of this thesis has direct practical implications for the conservation of Western European honey bees. Moreover, it can be used as an important case study to assess the genetic structure and investigate selection patterns in a complex subspecies.

References

Alberts, B. (2008). *Molecular biology of the cell: reference edition* (Vol. 1): Garland Science.

Arrese, E. L., & Soulages, J. L. (2010). Insect fat body: energy, metabolism, and regulation. *Annual Review of Entomology, 55*(1), 207-225. doi: 10.1146/annurev-ento-112408-085356

Bloch, G. (2010). The social clock of the honeybee. *Journal of Biological Rhythms, 25*(5), 307-317. doi: doi:10.1177/0748730410380149

Büchler, R., Costa, C., Hatjina, F., Andonov, S., Meixner, M. D., Le Conte, Y., Uzunov, A., Berg, S., Bienkowska, M., Bouga, M., Drazic, M., Dyrba, W., Kryger, P., Panasiuk, B., Pechhacker, H., Petrov, P., Kezić, N., Korpela, S., & Wilde, J. (2014). The influence of genetic origin and its interaction with environmental effects on the survival of *Apis mellifera L.* colonies in Europe. *Journal of Apicultural Research, 53*(2), 205-214. doi: 10.3896/IBRA.1.53.2.03

Dražić, M. M., Filipi, J., Prđun, S., Bubalo, D., Špehar, M., Cvitković, D., Kezić, D., Pechhacker, H., & Kezić, N. (2014). Colony development of two Carniolan genotypes (*Apis mellifera carnica*) in relation to environment. *Journal of Apicultural Research, 53*(2), 261-268. doi: 10.3896/ibra.1.53.2.07

Froy, O., & Miskin, R. (2007). The interrelations among feeding, circadian rhythms and ageing. *Progress in Neurobiology, 82*(3), 142-150. doi: https://doi.org/10.1016/j.pneurobio.2007.03.002

Hedin, L. E., Illergard, K., & Elofsson, A. (2011). An introduction to membrane proteins. *Journal of Proteome Research, 10*(8), 3324-3331. doi: 10.1021/pr200145a

Helm, B., Ben-Shlomo, R., Sheriff, M. J., Hut, R. A., Foster, R., Barnes, B. M., & Dominoni, D. (2013). Annual rhythms that underlie phenology: biological time-keeping meets environmental change. *Proceedings of the Royal Society B: Biological Sciences, 280*(1765). doi: 10.1098/rspb.2013.0016

López-Maury, L., Marguerat, S., & Bähler, J. (2008). Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nature Reviews Genetics, 9*(8), 583-593. doi: doi:10.1038/nrg2398

Louveaux, J., Albisetti, M., Delangue, M., & Theurkauff, M. (1966). Les modalités de L'adaptation des abeilles (*Apis mellifica L.*) au milieu naturel. *Ann. Abeille, 9*(4), 323-350.

Parejo, M., Wragg, D., Gauthier, L., Vignal, A., Neumann, P., & Neuditschko, M. (2016). Using Whole-Genome Sequence information to foster conservation efforts for the European Dark honey bee, *Apis mellifera mellifera*. *Frontiers in Ecology and Evolution, 4*, 140. doi: https://doi.org/10.3389/fevo.2016.00140

Peñuelas, J., & Filella, I. (2001). Responses to a warming World. *Science, 294*(5543), 793-795. doi: 10.1126/science.1066860

Reizer, J., Reizer, A., & Saier, M. H. (1994). A functional superfamily of sodium/solute symporters. *Biochimica et Biophysica Acta (BBA) - Reviews on Biomembranes, 1197*(2), 133-166. doi: http://dx.doi.org/10.1016/0304-4157(94)90003-5

Simpson, J. (1958). The problem of swarming in beekeeping practice. *Bee World, 39*(8), 193-202. doi: 10.1080/0005772x.1958.11095063

Sojo, V., Dessimoz, C., Pomiankowski, A., & Lane, N. (2016). Membrane proteins are dramatically less conserved than water-soluble proteins across the tree of life. *Molecular Biology and Evolution, 33*(11), 2874-2884. doi: 10.1093/molbev/msw164

Yerushalmi, S., & Green, R. M. (2009). Evidence for the adaptive significance of circadian rhythms. *Ecology Letters, 12*(9), 970-981. doi: 10.1111/j.1461-0248.2009.01343.x

Yurgel, M. E., Masek, P., DiAngelo, J., & Keene, A. C. (2015). Genetic dissection of sleep–metabolism interactions in the fruit fly. *Journal of Comparative Physiology A, 201*(9), 869-877. doi: 10.1007/s00359-014-0936-9

Chapter IX.

Supporting material and published papers

Supplementary Material for Chapter III

Supplementary Tables

Table Sup III-1 - Information content values of the initial 1183 SNP dataset estimated by the five selection methods (Weir & Cockerham's F_{ST}, Delta, informativeness (I_n), PCA and the F_{ST}-based outlier test) and for the four training datasets (I to IV). The SNPs are ordered from high to low information content. The top 48, 96, 144, 192 and 384 SNPs were included in the five reduced panels. SNPs marked with an asterisk (*) were excluded from the reduced panels because they were within a genetic distance < 1 cM of other informative SNPs **(available in a separate excel document).**

Table Sup III-2 - Information content values of the initial 1183 SNP dataset estimated by the five selection methods (Weir & Cockerham's F_{ST}, Delta, informativeness (I_n), PCA and the F_{ST}-based outlier test) and for the four training datasets (I to IV) **(available in a separate excel document).**

Table Sup III-3 - Admixture proportion estimates inferred from the five AIMs panels (48-, 96-, 144-, 192-, 384-AIMs) and the initial 1183 SNP dataset for the holdout set. The holdout set consisted of 34 pure (training set) and 43 reserved individuals of *A. m. mellifera* and all reference individuals of *A. m. ligustica* (17) and *A. m. carnica* (19). * Samples marked with an asterisk (*) are of *A. m. mellifera* from protected populations (pure breeding for conservation purposes; see Pinto *et al.* 2014 for details) **(available in a separate excel document).**

Table Sup III-4 – *P*-values of Mann-Whitney pairwise several-sample-test. Values obtained from comparing individual admixture proportions estimated with the five AIMs panels (48-, 96-, 144-, 192-, 384-AIMs) and the initial 1183 SNP dataset using the holdout set.

Panel	48-AIMs	96-AIMs	144-AIMs	192-AIMs	384-AIMs
96-AIMs	0.8837				
144-AIMs	0.9734	0.9199			
192-AIMs	0.9154	0.9983	0.9634		
384-AIMs	0.9065	0.9842	0.9327	0.9950	
1183 SNPs	0.9221	0.8200	0.9054	0.8450	0.8225

Table Sup III-5 - *P*-values of Mann-Whitney pairwise several-sample-test. Values obtained from comparing admixture proportions inferred from the five AIMs panels and the 1183 initial SNP dataset using the simulated set. The simulated set was generated with the program ONCOR (Kalinowski et al. 2007) using the function "simulate a single mixture". Ten populations, each with 100 genotypes, were simulated using different levels of C-lineage introgression (0, 1, 5, 10, 20, 30, 40, 50, 75, and 90%).

Panel	48-AIMs	96-AIMs	144-AIMs	192-AIMs	384-AIMs
96-AIMs	0.6580				
144-AIMs	0.8284	0.8207			
192-AIMs	0.9120	0.7392	0.9165		
384-AIMs	0.9742	0.6847	0.8625	0.9485	
1183 SNPs	0.4588	0.2313	0.3355	0.3943	0.4442

Supplementary Figures

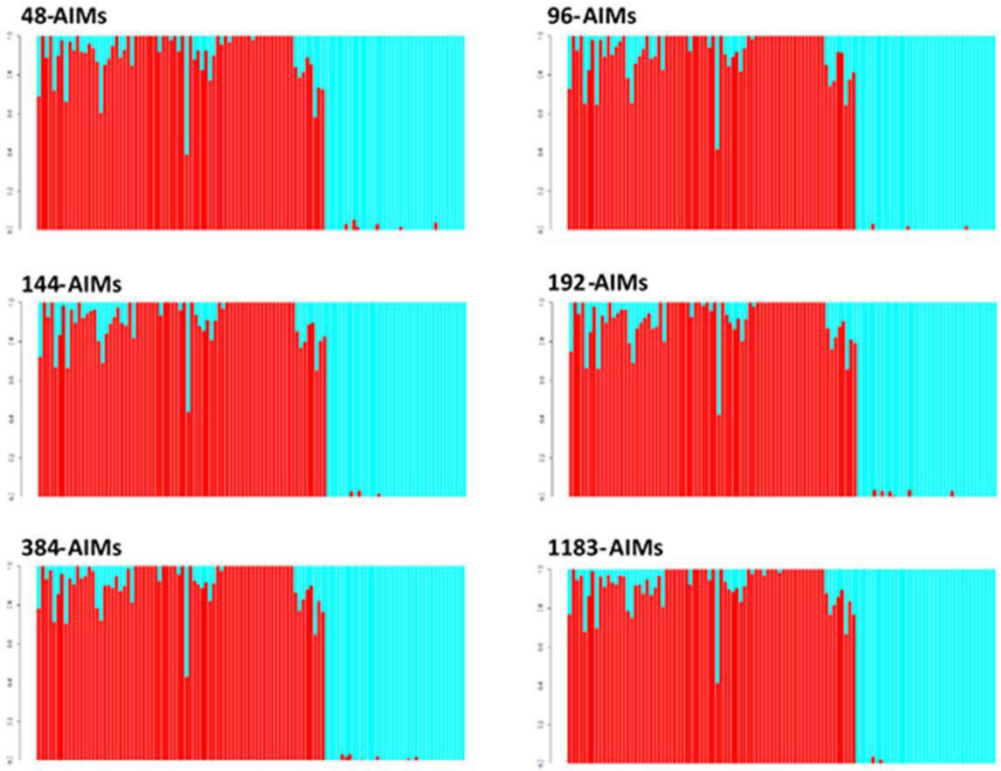


Figure Sup III-1 - Ancestry estimates. Global estimates (y-axis), for the 113 individuals of the holdout set (x-axis), inferred from the five AIMs panels (48-, 96-, 144-, 192-, 384-AIMs) and the initial 1183 SNP dataset using the model-based approach implemented in the ADMIXTURE software. Results are shown for the optimal *K*=2, which distinguishes the M (red) and C (cyan) evolutionary lineages of *A. mellifera*.

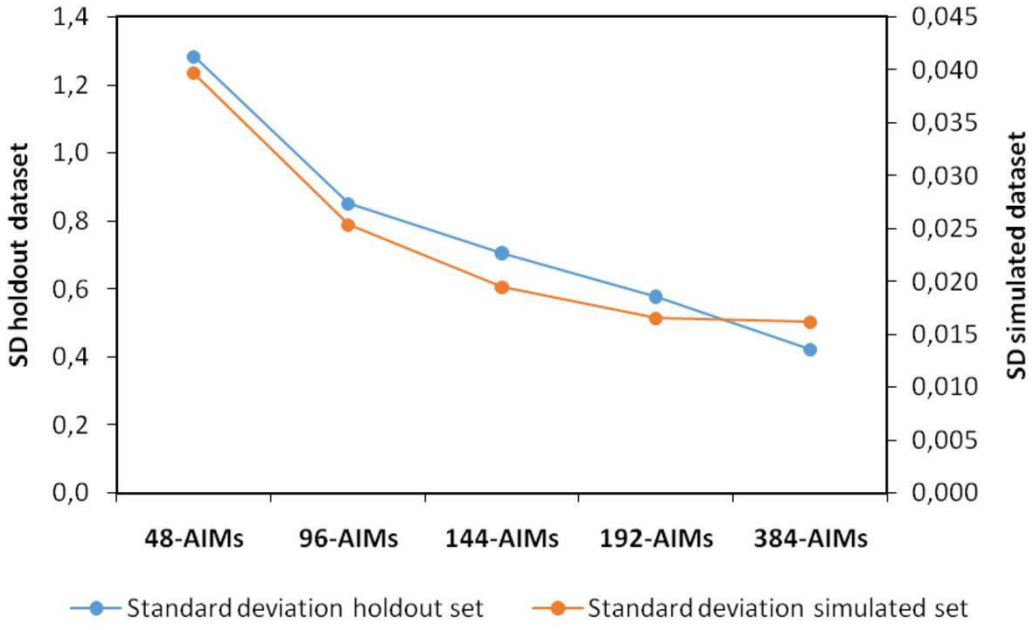


Figure Sup III-2 - Standard deviation (SD) of admixture proportions. Precision estimates obtained using the SD of the differences between admixture proportions inferred from the initial 1183 SNP dataset and the five AIMs panels (48-, 96-, 144-, 192-, 384-AIMs) using the holdout (blue line) and simulated (orange line) sets.

Supplementary Material for Chapter IV

Supplementary Tables (available in a separate excel document)

Table Sup IV-1 - Information for the 127 highly-informative SNPs selected by the Assay Design software, including flanking regions and [SNP], primers used for genotyping in the Agena Bioscience's MassARRAY® MALDI-TOF platform, missing data before (*) and after (**) the quality control step, and statistics (F_{ST}, Delta, In, PCA, and F_{ST} outlier test) used by Muñoz, *et al.* 2015 to rank SNPs by information content. SNPs with >20% of missing data and that were heterozygous (Ho) in >10% of the haploid individuals were discarded. The 29 SNPs with 100% success rate in the sensitivity test are marked in bold.

Table Sup IV-2 - Information for the 573 samples genotyped for the four SNP assays in the Agena Bioscience's MassARRAY® MALDI-TOF platform. The percentages of missing data, calculated before and after the quality control step, are identified by the symbols † and ‡, respectively. Samples with >30% of missing data were excluded from further analysis. Samples marked by the symbol * were also genotyped with the GoldenGate® Assay in the Illumina's BeadArray platform Pinto, *et al.* 2014. Samples marked by ** were genotyped concurrently in the MassARRAY® MALDI-TOF platform, in the BeadArray platform, and in the Illumina's HiSeq 2500 platform (D.H./M.A.P., unpublished data; Parejo, *et al.* 2016. Subspecies names were inferred from location of sampled colonies and native ranges reported by Ruttner (1988).

Table Sup IV-3 - Filtering criteria and number of SNPs filtered out.

Table Sup IV-4 - Number of SNPs per linkage group contained in the four SNPs assays (M1, M2, M3, M4) before (on the left) and after (on the right) the quality control step.

Table Sup IV-5 - Q-values inferred from the four assays (117 SNPs) by STRUCTURE and ADMIXTURE for 112 drones, of which 96 were genotyped in the MassARRAY® MALDI-TOF platform and 16 (9 *A. m. carnica* and 7 *A. m. ligustica*) were previously genotyped using the GoldenGate® Assay in the Illumina's BeadArray platform Pinto, *et al.* 2014 and added to the dataset for a better representation of lineage C.

Table Sup IV-6 - Q-values inferred from the four assays (117 SNPs) by ADMIXTURE for haploid drones (N=112), diploid workers (N=112), and their combination (N=224). Of the 112 drones, 96 were genotyped in the MassARRAY® MALDI-TOF platform and 16 (9 *A. m. carnica* and 7 *A. m. ligustica*) were previously genotyped using the GoldenGate® Assay in the Illumina's BeadArray platform Pinto, *et al.* 2014 and added to the dataset for a better representation of lineage C.

Table Sup IV-7 - Q-values inferred from genome-wide SNPs, identified in whole genome (WG) sequences, and from the four SNP assays (M1, M2, M3, M4) for 38 individuals. Of these, 32 individuals were genotyped in the MassARRAY® MALDI-TOF platform and six were previously genotyped using the GoldenGate® Assay in the Illumina's BeadArray platform Pinto, *et al.* (2014). Subspecies names were inferred from location of sampled colonies and native ranges reported by Ruttner (1988).

Table Sup IV-8 - Q-values inferred from the four SNP assays (M1, M2, M3, M4) for the offspring of the different crosses used in the validation test. The thresholds for the expected Q-values were >0.95 for *A. m. carnica*, <0.05 for *A. m. mellifera* and 0.50 for the F1 hybrids.

Table Sup IV-9 - Q-values inferred from the expected genotype (as determined by the SNPs called individually in the drone samples 2683 and 2802) and the observed genotypes obtained for the DNA pools (dilution ratios of 10:20, 5:40, 2:20, 1:20, 2:20, 0.5:20) using the four SNP assays (117 SNPs), the two best assays M1+M3 (62 SNPs) and the 29 cross-detected SNPs.

Table Sup IV-10 - Information on workers (and their progenitors) used to validate the four SNPs assays and to construct the tissue pools for assessing the sensitivity of the MassARRAY® MALDI-TOF genotyping system. The pools were constructed by mixing varying ratios of half thoraces. The DNA concentrations obtained from single workers and pools of workers were measured using NanoDropTM.

Table Sup IV-11 - Number of SNPs accurately called in the 22 tissue pools.

Table Sup IV-12 - Number of SNPs accurately called and miscalled for each of the 22 tissue pools.

Table Sup IV-13 - Q-values inferred from the expected and observed genotypes obtained for the 22 tissue pools called using the four assays (117 SNPs), the two best assays M1+M3 (62 SNPs), and the 29 cross-detected SNPs identified in the pooled-DNA experiment. The expected genotypes were inferred for the 117 SNP loci from the calls obtained for the single workers.

Table Sup IV-14 - Q-values (individual values, mean, and standard deviation) inferred from the four assays (117 SNPs) by ADMIXTURE for colonies of varying ancestry sampled across Europe, as shown in Fig. 1. Each sample represents a single individual and colony. Samples of A*. m. mellifera* were collected from colonies in protected and unprotected apiaries. A total of nine *A. m. carnica* and seven *A. m. ligustica* samples, previously genotyped using the GoldenGate® Assay in the Illumina's BeadArray platform Pinto, *et al.* (2014), was added to the dataset for a better representation of lineage C. Subspecies names were inferred from location of

sampled colonies and native ranges reported by Ruttner (1988), except for two samples from Scotland that were purchased to an *A. m. carnica* breeder.

Table Sup IV-15 - Q-values (individual values, mean, and standard deviation) inferred from the four assays (117 SNPs) by ADMIXTURE for *A. m. mellifera* colonies sampled (pools and single individuals) in protected and unprotected apiaries from the United Kingdom and Switzerland. Q-values for colonies of *A. m. carnica* and Buckfast are also shown. Subspecies names were inferred from location of sampled colonies and native ranges reported by Ruttner (1988), except for one sample from Scotland that was purchased to an *A. m. carnica* breeder.

Table Sup IV-16 - Genotypes of the 566 quality-proved samples for 117 SNP loci (four assays) typed in the MassARRAY® MALDI-TOF platform. A total of nine *A. m. carnica* and seven *A. m. ligustica* samples, previously genotyped using the GoldenGate® Assay in the Illumina's BeadArray platform Pinto, *et al.* (2014), was added to the dataset for a better representation of lineage C. The 16 C-lineage colony samples are marked in bold. Subspecies names were inferred from location of sampled colonies and native ranges reported by Ruttner (1988).

Table Sup IV-17 - Combinations of workers of varying ancestry used to construct the 22 tissue pools for assessing the sensitivity of the MassARRAY® MALDI-TOF genotyping system.

Supplementary Material for Chapter V

Supplementary Tables (available in a separate excel document)

Table Sup V-1 - Correlation between the extracted environmental variables.precipitation (prec), minimum temperature (tmin), mean temperature (tmean), maximum temperature (tmax), cloud cover (cld), relative humidity (rh), and insolation (ins). Arabic numerals from 1 to 12 on front of each environmental variable designate the month for which the variable was obtained.

Table Sup V-2 - The 13 climatic variables retained for analysis and their correlated variables. The r (correlation) values are shown within parenthesis. The uncorrelated variables were: longitude (long), latitude (lat), altitude (alt), precipitation in January (prec1), precipitation in May (prec5), precipitation in August (prec8), minimum temperature in January (tmin1), minimum temperature in June (tmin6), cloud cover in April (cld4), cloud cover in July (cld7), relative humidity in January (rh1), relative humidity in March (rh3), relative humidity in June (rh6), and insolation in April (ins4).

Table Sup V-3 - Mean coverage of each sample and sampling site codes.

Table Sup V-4 - Filtering criteria and percentage of SNPs that were filtered out.

Table Sup V-5 - Distribution of the 1,289,449 SNPs across genomic regions.

Table Sup V-6 - Genomic information and statistics for the top-ranked 4,290 models (P-value<0.0001) obtained by Samβada. The uncorrelated variables were: longitude (long), latitude (lat), altitude (alt), precipitation in January (prec1), precipitation in May (prec5), precipitation in August (prec8), minimum temperature in January (tmin1), minimum temperature in June (Tmin6), cloud cover in April (cld4), cloud cover in July (cld7), relative humidity in January (rh1), relative humidity in March (rh3), relative humidity in June (rh6), insolation in April (ins4), and land cover. For each SNP we have the corresponding gene and functional state, for the coding sequences (cds) inside of the parenthesis is the information about if the SNP is in a synonymous (syn) or non-synonymous (non) positions. For the intronic and intergenic SNP inside of the parentheses is the distance in bp from the nearest coding region.

Table Sup V-7 - Genomic information and statistics for the SNPs identified by LFMM with a FDR < 0.05 and associated environmental variables. The uncorrelated variables were: longitude (long), latitude (lat), altitude (alt), precipitation in January (prec1), precipitation in May (prec5), precipitation in August (prec8), minimum temperature in January (tmin1), minimum temperature in June (tmin6), cloud cover in April (cld4), cloud cover in July (cld7), relative humidity in January (rh1), relative humidity in March (rh3), relative humidity in June (rh6), insolation in April (ins4), and land cover. For each SNP we have the corresponding gene and functional state, for the coding sequences (cds) inside of the parenthesis is the information about if the SNP is in a synonymous (syn) or non-synonymous (non) positions. For the intronic and intergenic SNP inside of the parentheses is the distance in bp from the nearest coding region.

Table Sup V-8 - Candidate genes and intergenic regions detected by at least two methods. [1]For each genes there is information if SNPs lie in the genic region (exon, introns and UTRs) or in the intergenic region near the corresponding gene. In the cases where there are SNIPs in the genic and intergenic region the number of SNIPs in each class is inside of parenthesis [2]Orthologs in *D. melanogaster* (coded by FBgnxxxxxxx); genes with no orthologs in *D. melanogaster* were classified into four groups: orphan, Apoidea, Hymenoptera or higher taxonomic categories using OrthoDB V.8. 2 Number of SNPs detected by at least two methods are indicated within parenthesis. [4]Number of SNPs and associated environmental variables are indicated within parenthesis. [5,6,7] Functional annotations obtained from FLYBASE acceded in June 2016. [8] Genes detected by other studies. *Genes that contain SNPs detected by at least three methods. **Genes that contain SNPs detected by four methods.

Table Sup V-9- The outlier SNPs detected by at least two selection methods and the correspondent |iHs| and –log10 (q-value) obtained with PCAdapt and associated environmental variables detected by both GEA methods (int). For each SNP we have the corresponding gene and functional state, for the coding sequences (cds) inside of the parenthesis is the information about if the SNP is in a synonymous (syn) or non-synonymous (non) positions. For the intronic and intergenic SNP inside of the parentheses is the distance in bp from the nearest coding region.

Table Sup V-10- Number of shared outlier SNPs among the environmental variables. The diagonal elements are the number of SNPs concurrently detected by both GEA methods, below and above the diagonal are the numbers and percentage of shared SNPs, respectively.

Table Sup V-11- Iberian honeybee candidate genes detected by PCAdapt and corresponding number of SNPs harbored by those genes. Orthologs in *D. melanogaster* (coded by FBgnxxxxxxx); genes with no orthologs in *D. melanogaster* were classified into four groups: orphan, Apoidea, Hymenoptera or higher taxonomic categories using OrthoDB V.8. The 22 genes marked in bold were cross-validated by LFMM and Samβada.

Table Sup V-12- Candidate genes, detected by at least two methods, containing SNPs in non-synonymous positions.

Table Sup V-13- RMSD, ΔΔG and total energy after minimization values of reference BeeBase protein with 3D predicted structure. Iberian honeybee individuals harboring sequence variants identical to the BeeBase reference are marked by an + Variants were named by upper case letter starting with A as the most frequent and following until the less frequent.

Table Sup V-14- Candidate genes listed by biological process chosen among the 305 genes. The number of genes is indicated within parenthesis.

Table Sup V-15- Significantly enriched (P-value<0.05) annotation terms for the genes harbouring SNPs detected by at least two methods. Similar annotation terms were clustered together.

Supplementary Figures

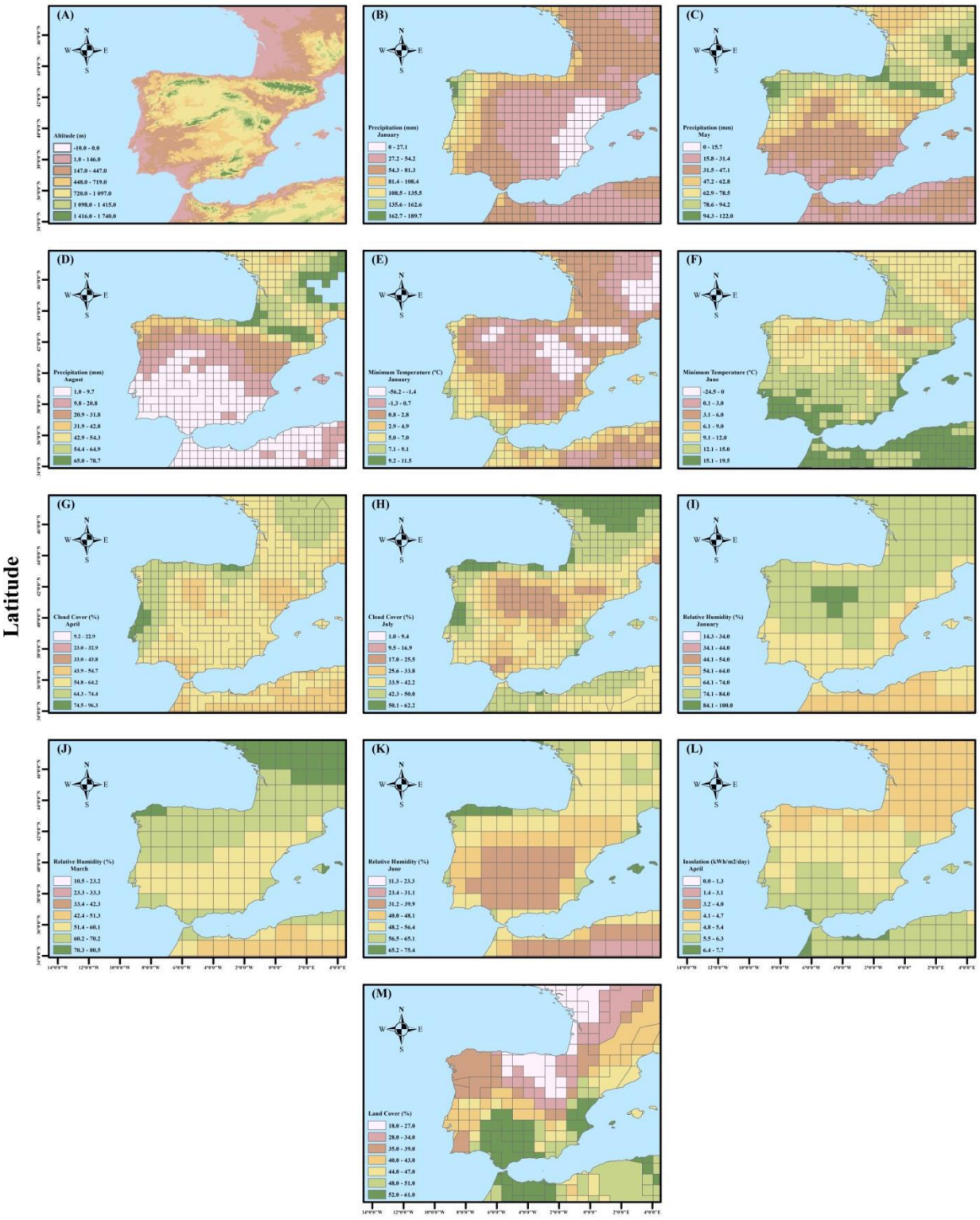


Figure Sup V-1- Maps showing the distribution of the 13 climatic variables used in the analyses. (A) altitude (alt), (B) precipitation in January (prec1), (C) precipitation in May (prec5), (D) precipitation in August (prec8), (E) minimum temperature in January (tmn1), (F) minimum temperature in June (tmn6), (G) cloud cover in April (cld4), (H) cloud cover in July (cld7), (I) relative humidity in January (rh1), (J) relative humidity in March (rh3), (K) relative humidity in June (rh6), (L) insolation in April (ins4), and (M) land cover.

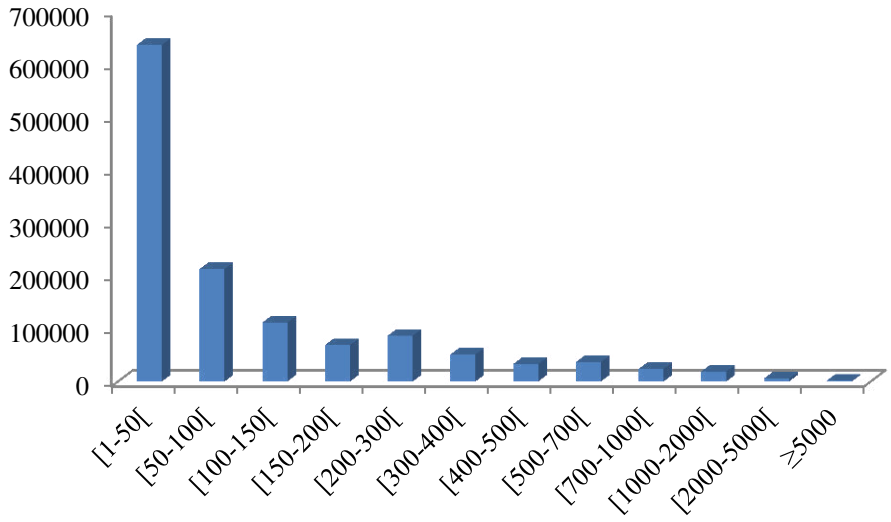


Figure Sup V-2. Distribution of the distance between the 1,289,449 SNPs across genomic regions. The average physical distance between SNPs was 170.262 bp varying between 1 bp and 136,266 bp.

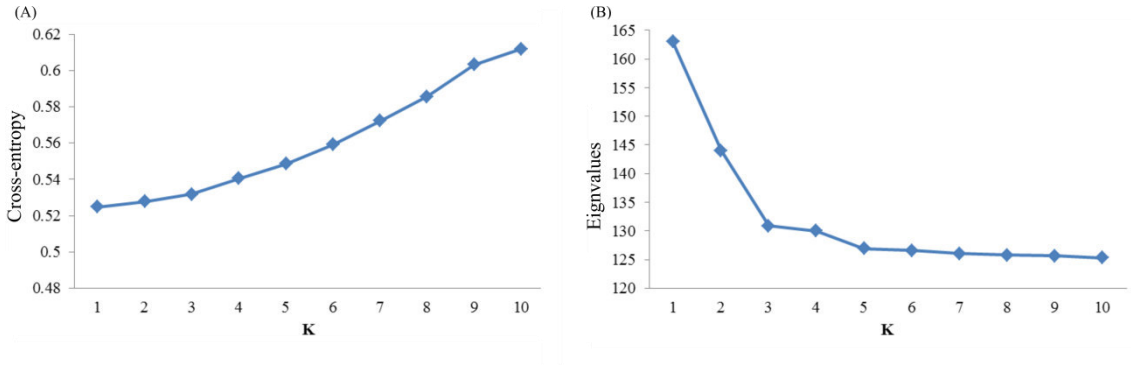


Figure Sup V-3. Graphical display of the two methods (cross-entropy and eigenvalue) used to predict the optimal K in the analysis of population structure

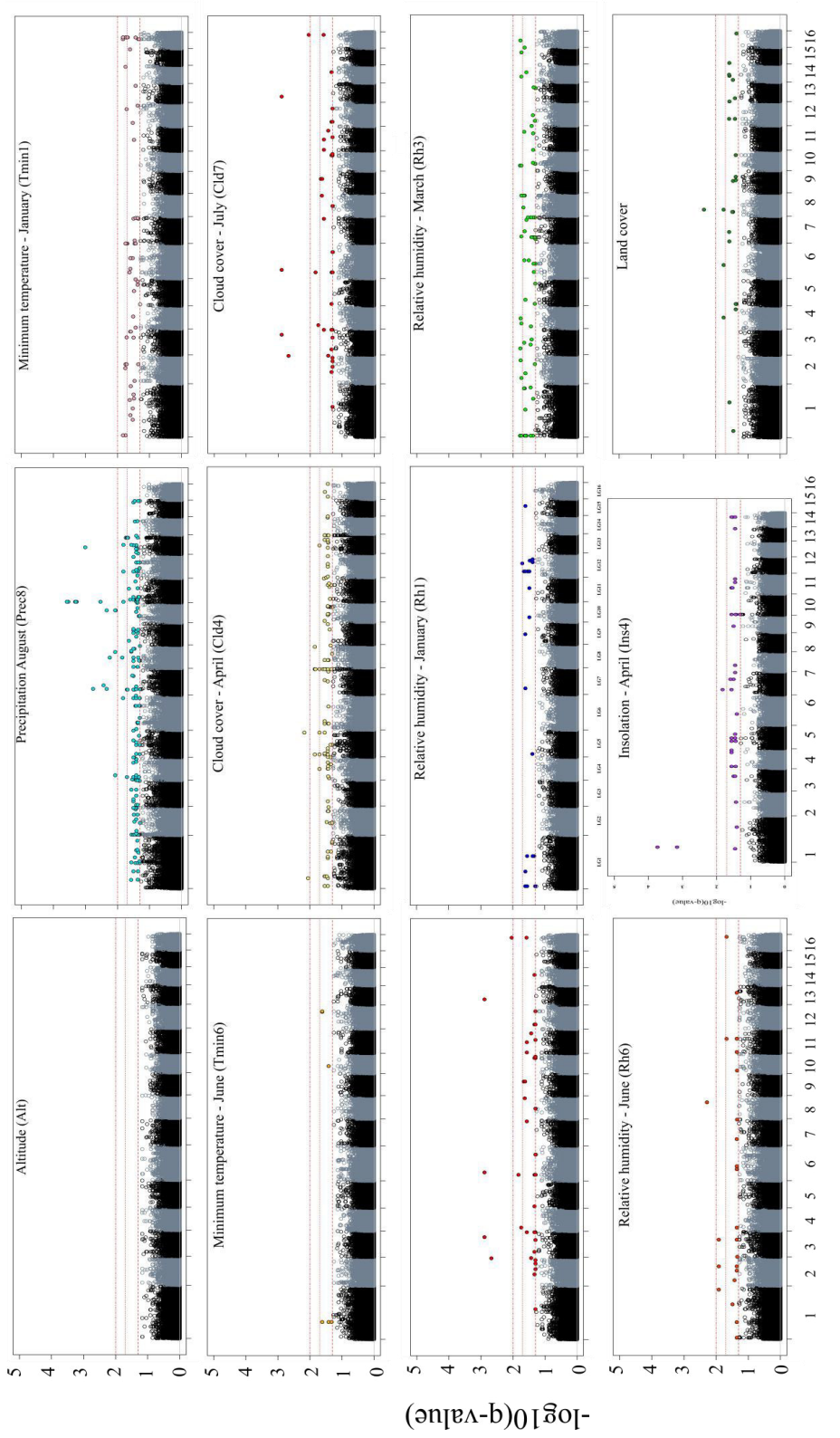


Figure Sup V-4. Manhattan plots representing the genome-wide distribution of significance values –log10 (*q*value) obtained by the genetic-environment association approach LFMM for 13 environmental variables. The red lines indicate FDR values of 0.05, 0.02 and 0.01. The variables lat, long, and cld4 exhibited 385, 113, and 197 significant associations, respectively (see Table 1 and supplementary table S7, Supplementary Material online)

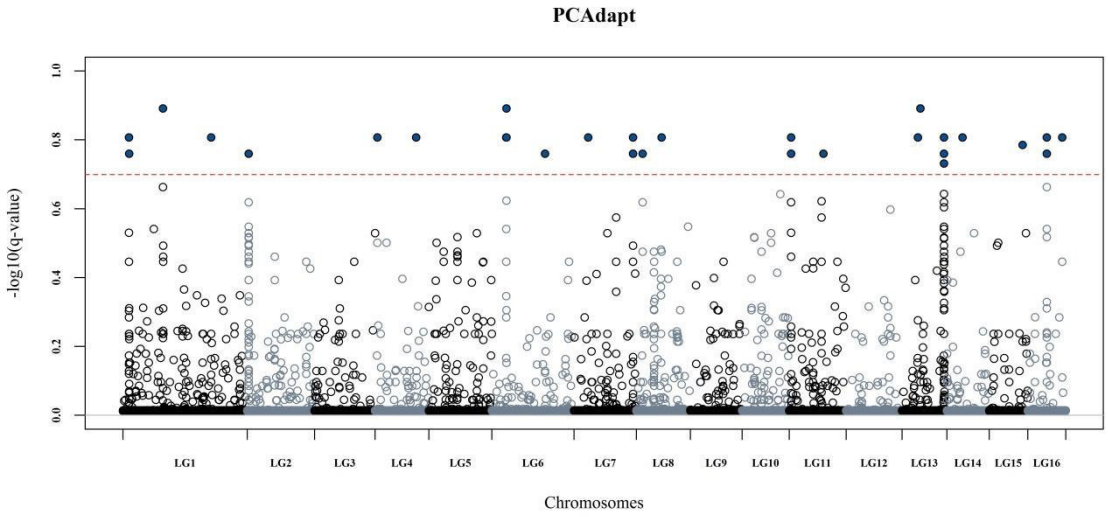
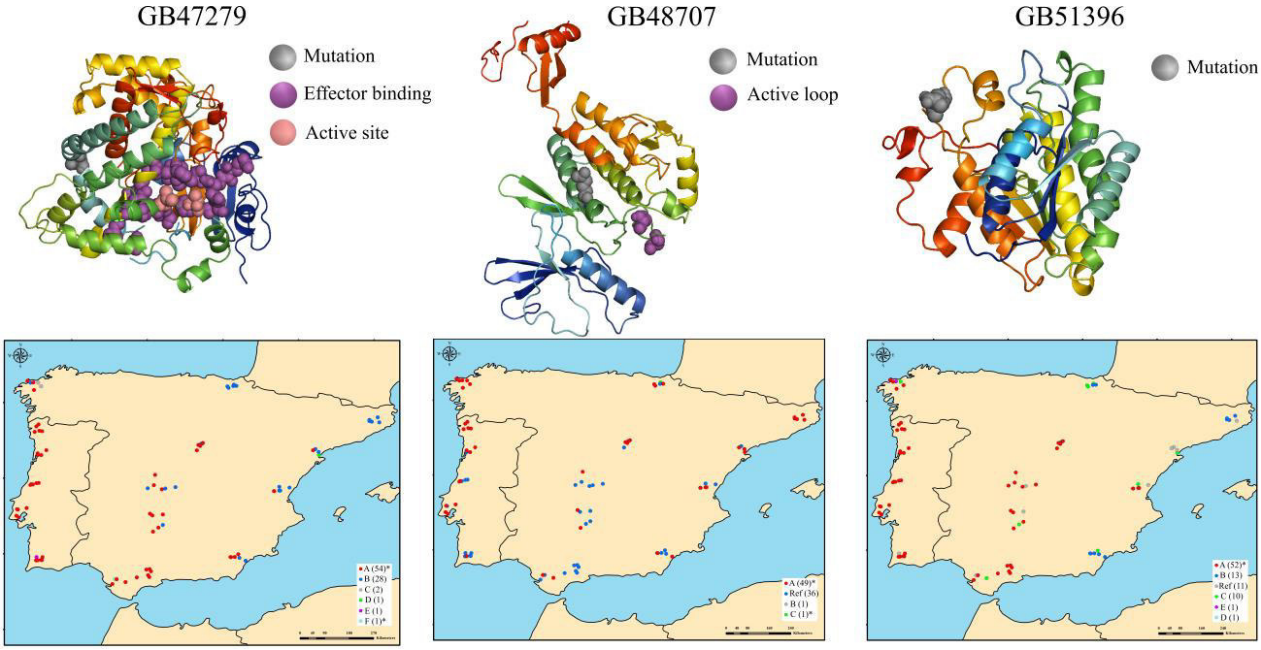


Figure Sup V-5- Manhattan plot representing the genome-wide distribution of significance values –log10 (*q*-value) obtained by PCAdapt. The red line indicates FDR<0.2.

Figure Sup V-6- Predicted protein structures of three genes harbouring non-synonymous outlier SNPs



detected by three genome-wide methods. The structures were predicted by Pymol considering the BeeBase reference amino acid sequences. The grey spheres represent the position and altered amino acids. The asterisk represents the variants carrying the SNP with signs of selection. The maps show how the different protein variants gather in space. Numbers in parentheses show the number of individuals.

Supplementary Material for Chapter VI

Supplementary Material and Methods

SEQUENCING, MAPPING, QUALITY CONTROL, AND VARIANT CALLING

Paired-end sequencing libraries (2x100/150bp) from whole-genomic DNA of single drone samples were prepared following the manufacturer's protocol (TruSeq Nano Kit) and sequenced on an Ilumina HiSeq aiming at a depth of coverage of 10X. The choice of drones in this re-sequencing project stems from their haploidy, which enabled to obtain phased data and to confidently identify SNPs with less coverage than diploid individuals (Wragg *et al.*, 2016). Sequencing reads were mapped against the reference genome (Amel4.5) using bwa mem 0.7.10 (Li & Durbin, 2009) and PCR duplicates marked using PICARD 1.80 (http://picard.sourceforge.net/). GATK 3.3.0 (McKenna *et al.,* 2010; Van der Auwera *et al.,* 2013) was used to improve mapping quality and read realignment around the indels. The depth of coverage (DP) statistics was calculated for each sample (bam file) with BEDtools (Quinlan & Hall, 2010).

SNP calling was performed using a two-step process following Wragg *et al.* (2016). Briefly, SNP variants were first identified for each sample separately applying three different variant calling tools: GATK's UnifiedGenotyper (Van der Auwera *et al.*, 2013), SAMtools' mpileup 1.1 (Li et al., 2009) and PLATYPUS 0.8.1 (Rimmer *et al.*, 2014). Variants were then filtered according to base quality (BQ) score ≥20 and mapping quality ≥30. Calls from UnifiedGenotyper were additionally filtered for a maximum number of two alternate alleles, genotype quality ≥30, quality by depth ≥2, and Fisher strand ≤60. After this quality control step, the three call sets were combined using BAYSIC (Cantarel *et al.*, 2014), for the datasets of Wragg *et al.* (2017) and Parejo *et al.* (2016), and BCFtools (Li, 2011), for the dataset of Henriques (DH and MAP, unpublished data). Variant calling statistics was calculated for each sample with VCFtools (Danecek *et al.,* 2011). The single-sample variant calling files (VCFs) were then merged keeping only SNP variants, which were filtered

on 9≤DP≤3X mean DP to generate a set of master sites, mapped to chromosomes 1 to 16, using BCFtools. All individuals were re-genotyped with BQ≥20 at these master sites resulting in a multi-sample VCF comprising all samples.

As a quality control step, SNPs were filtered using PLINK 1.9 (Chang et al., 2015) excluding SNPs with minor allele frequency (MAF) <0.01 and genotyping call rate <0.9. Finally, functional SNP annotation was performed using SNPeff 4.1a (Cingolani et al., 2012) and the reference genome amel4.5.

SAMPLE SELECTION

A pre-requisite for developing reliable reduced SNP panels for diverse applications is to have reference samples that truly represent the original populations. For example, individuals of A. m. carnica showing M-lineage introgression must be removed from a reference population. With this in mind, we started with an initial whole-genome dataset of 313 individuals (Parejo *et al.,* 2016; Wragg *et al.,* 2017; D. Henriques, unpublished data), which included A. m. iberiensis (N=117), *A. m. ligustica* (N=34), *A. m. carnica* (N=37), *A. m. mellifera* (N=111), and the commercial Buckfast breed (N=14), to assess individuals' purity using ADMIXTURE v1.3.0 (Alexander *et al.,* 2009). The ancestry proportions of these samples were inferred using the model-based clustering for K=1 to 5 ancestral populations and the default termination criterion set to stop when the log-likelihood increases by less than 0.0001 between independent runs. At the optimal K=2, all *A. m. iberiensis* (N=117) were selected, as none of the samples was admixed, and a total of 59 C-lineage individuals (28 *A. m. carnica* and 31 *A. m. ligustica*) with <5% M-lineage ancestry were included in the final dataset. An additional filtering step of MAF <0.05 was applied on the final dataset (N=176) resulting in a whole-genome dataset of 2,366,382 SNPs.

REFERENCES

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Research, 19(9), 1655-1664. doi: 10.1101/gr.094052.109

Cantarel, B. L., Weaver, D., McNeill, N., Zhang, J., Mackey, A. J., & Reese, J. (2014). BAYSIC: A Bayesian method for combining sets of genome variants with improved specificity and sensitivity. BMC Bioinformatics, 15(1). doi: 10.1186/1471-2105-15-104

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience, 4, 7. doi: 10.1186/s13742-015-0047-8

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin), 6(2), 80-92. doi: 10.4161/fly.19695

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. Bioinformatics, 27(15), 2156-2158. doi: 10.1093/bioinformatics/btr330

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics, 27(21), 2987-2993. doi: 10.1093/bioinformatics/btr509

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14), 1754-1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078-2079. doi: 10.1093/bioinformatics/btp352

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research, 20(9), 1297-1303. doi: 10.1101/gr.107524.110

Parejo, M., Wragg, D., Gauthier, L., Vignal, A., Neumann, P., & Neuditschko, M. (2016). Using whole-genome sequence information to foster conservation efforts for the European Dark honey bee, *Apis mellifera mellifera*. Frontiers in Ecology and Evolution, 4, 140.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6), 841-842. doi: 10.1093/bioinformatics/btq033

Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R. F., Wilkie, A. O. M., McVean, G., & Lunter, G. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nature Genetics, 46(8), 912-918. doi: 10.1038/ng.3036

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Current Protocols in Bioinformatics(SUPL.43). doi: 10.1002/0471250953.bi1110s43

Wragg, D., Marti-Marimon, M., Basso, B., Bidanel, J.-P., Labarthe, E., Bouchez, O., Le Conte, Y., & Vignal, A. (2016). Whole-genome resequencing of honeybee drones to detect genomic selection in a population managed for royal jelly. 6, 27168.

Wragg, D., Basso, B., Beguin, M., Canale-Tabet, K., Costa, C., Gregorc, A., ... Vignal, A. (2017). Understanding the French honeybee populations by whole genome sequencing of haploid drones. Manuscript in preparation.

Supplementary Tables (available in a separate excel document)

Table Sup VI-1- Number of variable nucleotides on either side of the 250 bp flanking sequences of the fixed SNPs. Marked in bold are the SNPs used for final assay selection.

Table Sup VI-2- Sample origin, coverage statistics and number of variants.

Table Sup VI-3- Genomic information for the 18,272 fixed SNPs (FST=1) between *A. m. iberiensis* and C-lineage honeybees.

Table Sup VI-4- Distribution of SNPs across the 16 chromosomes obtained in the different phases of the assay design.

Table Sup VI-5- Distribution of fixed SNPs by sequence ontology terms.

Table Sup VI-6- Number of fixed SNPs distributed in 1,347 genic regions (±5 kb around coding sequences).

Table Sup VI-7- Significantly enriched gene ontology (GO) terms (P-value < 0.05) for the 1,347 genes harbouring fixed SNPs.

Table Sup VI-8- Performance comparison of reduced (M1-M4) and random (R1-R4) SNP assays in estimating C-lineage proportions (Q-value) in *A. m. iberiensis* holdout and simulated datasets.

Supplementary Figures

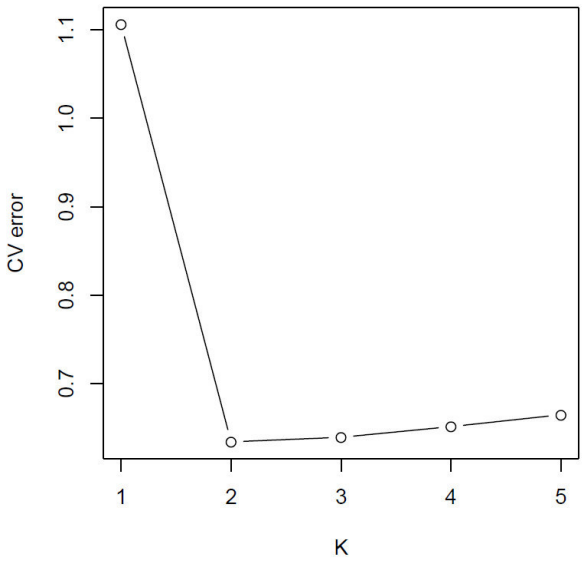


Figure Sup VI-1- Cross validation (CV) errors for K=1 to 5. CV error is lowest at K=2, suggesting optimal clustering with two ancestral populations.

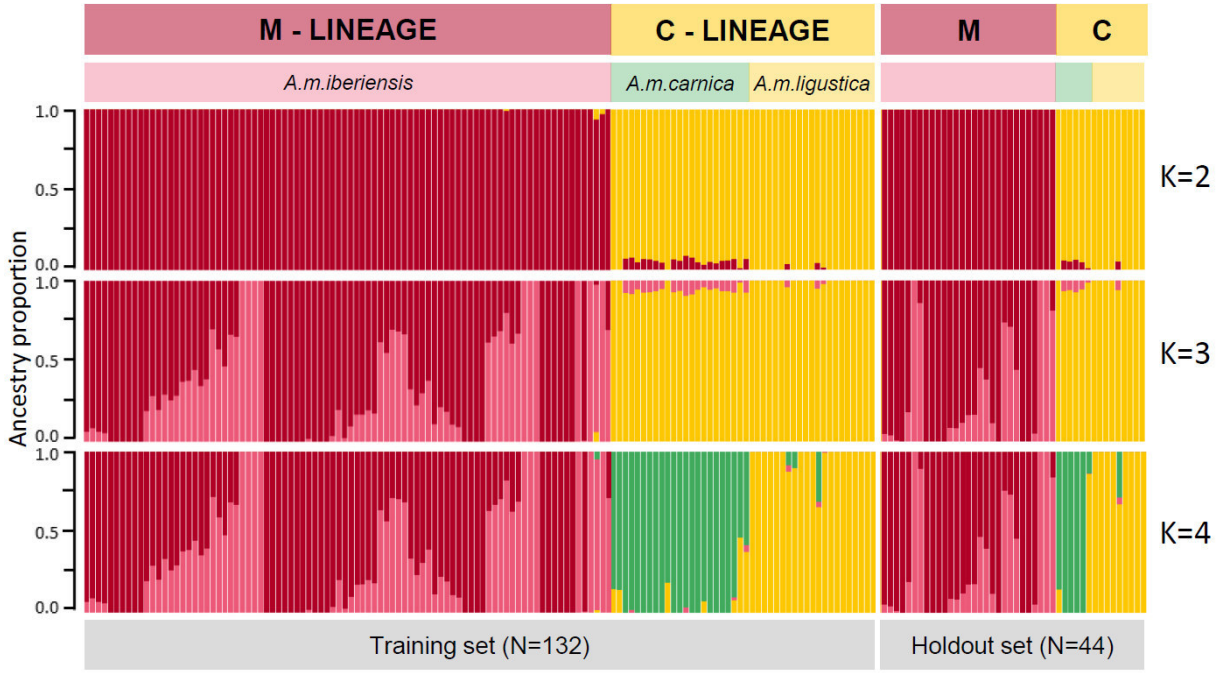


Figure Sup VI-2- Ancestry proportions (y-axis) from K=2 to 4 clusters. Each individual is represented by a vertical bar and samples are grouped according to subspecies and sample (training and holdout dataset). Each colour represents one cluster and individuals are coloured according to the proportion of the genome that was derived from each cluster. At K=2, the individuals are separated in accordance with the evolutionary lineages M and C. At K=3, the substructure within *A. m. iberiensis* is revealed, in line with the geographical cline of this subspecies in Iberia (Chávez-Galarza et al., 2015). At K=3, C-lineage individuals are subdivided into *A. m. carnica* and *A. m. ligustica*.

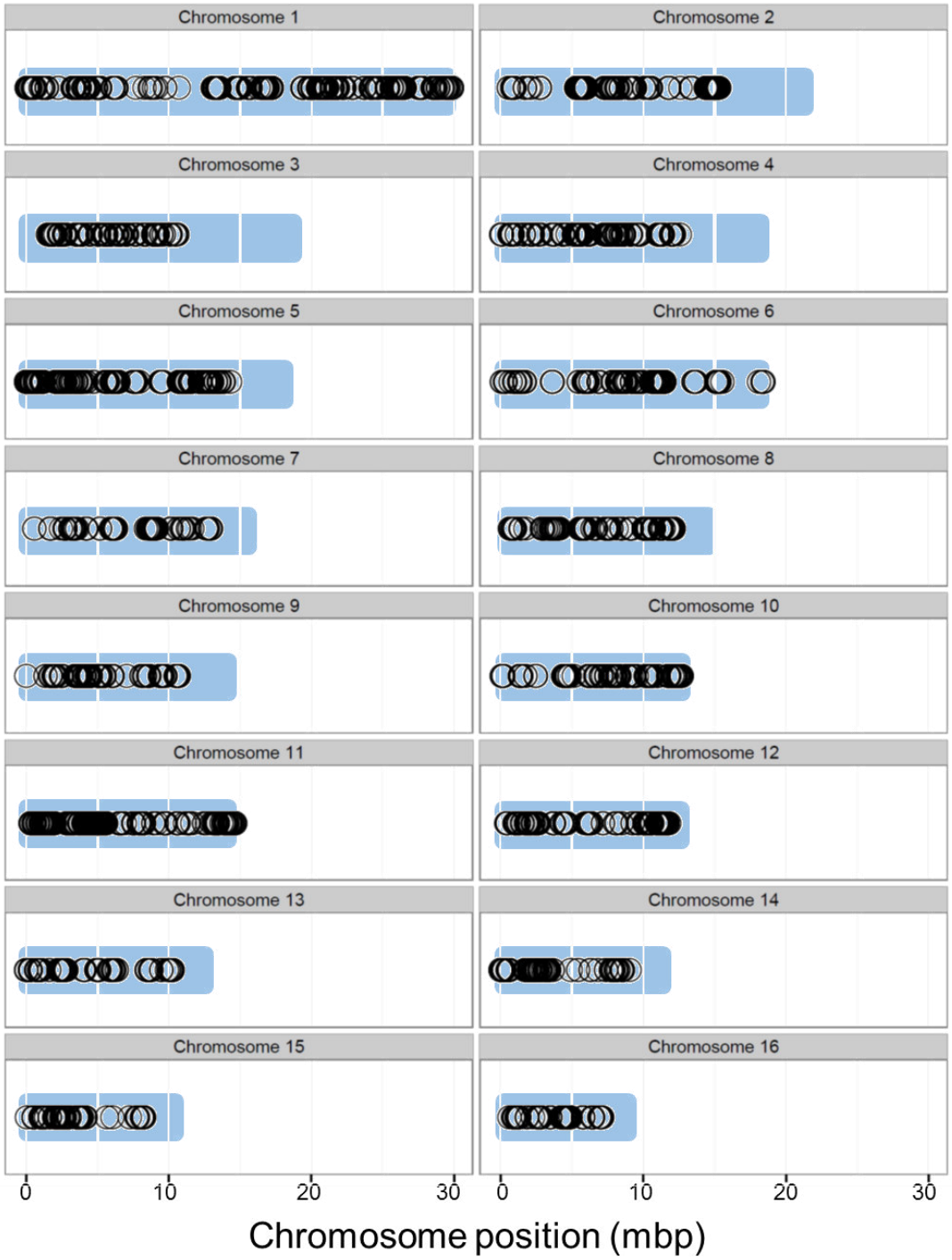


Figure Sup VI-3- Positions of fixed SNPs between *A. m. iberiensis* and C-lineage subspecies (*A. m. carnica* and *A. m. ligustica*) along the 16 honeybee chromosomes (see Table S2 for more detailed information).

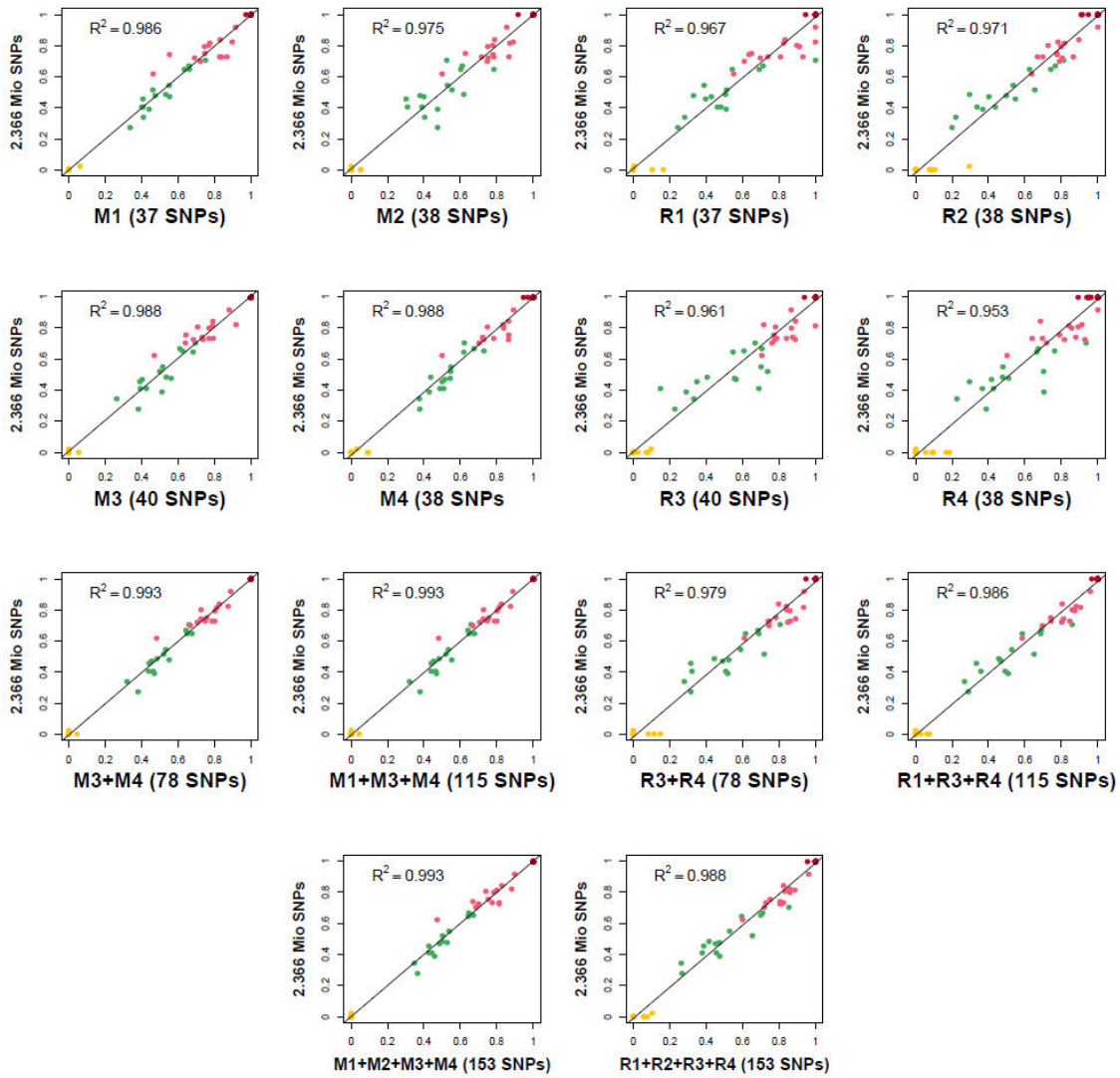


Figure Sup VI-4- Linear regression analyses for the ultra-low-density (M1, M2, M3, M4) and random (R1, R2, R3, R4) SNP assays (single or combined) against the whole-genome SNPs.

Supplementary Material for Chapter VII

Supplementary Tables (available in a separate excel document)

Table Sup VII-1. Sample information. Mean coverage of the mitogenome for each sample. Each haplotype was nominated using the nomenclature system revised by Chávez-Galarza et al. (2017) for the tRNAleu-cox2 intergenic region.

Table Sup VII-2-. Filtering criteria and number of SNPs that were filtered out using an initial data set of 123 individuals representing seven honey bee subspecies.

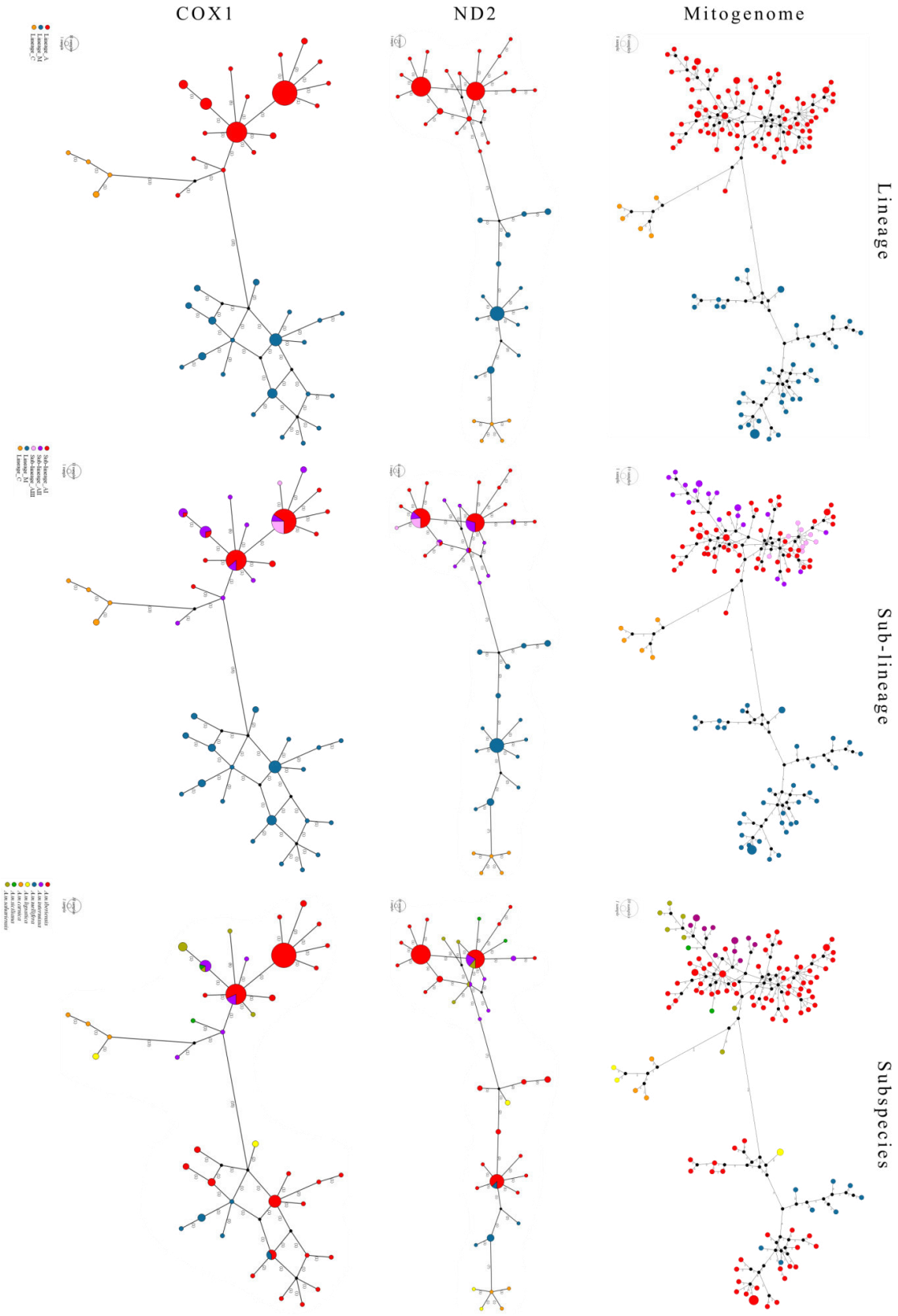
Table Sup VII-3- General information about the 17 datasets. All the protein coding and rRNA genes consist on different datasets referred by their common name. The dataset named as protein coding is the result of all the genes concatenated. Finally, the mitogenome dataset consists on nearly complete mtDNA, since for most of the individuals the tRNAleu-cox2 intergenic region is not complete. The position of protein and rRNA genes in relation to the reference genome is indicated in Start and Stop columns. Strand indicates the coding strand (- antisense + sense). Some information it is not applied (NA) to some databases.

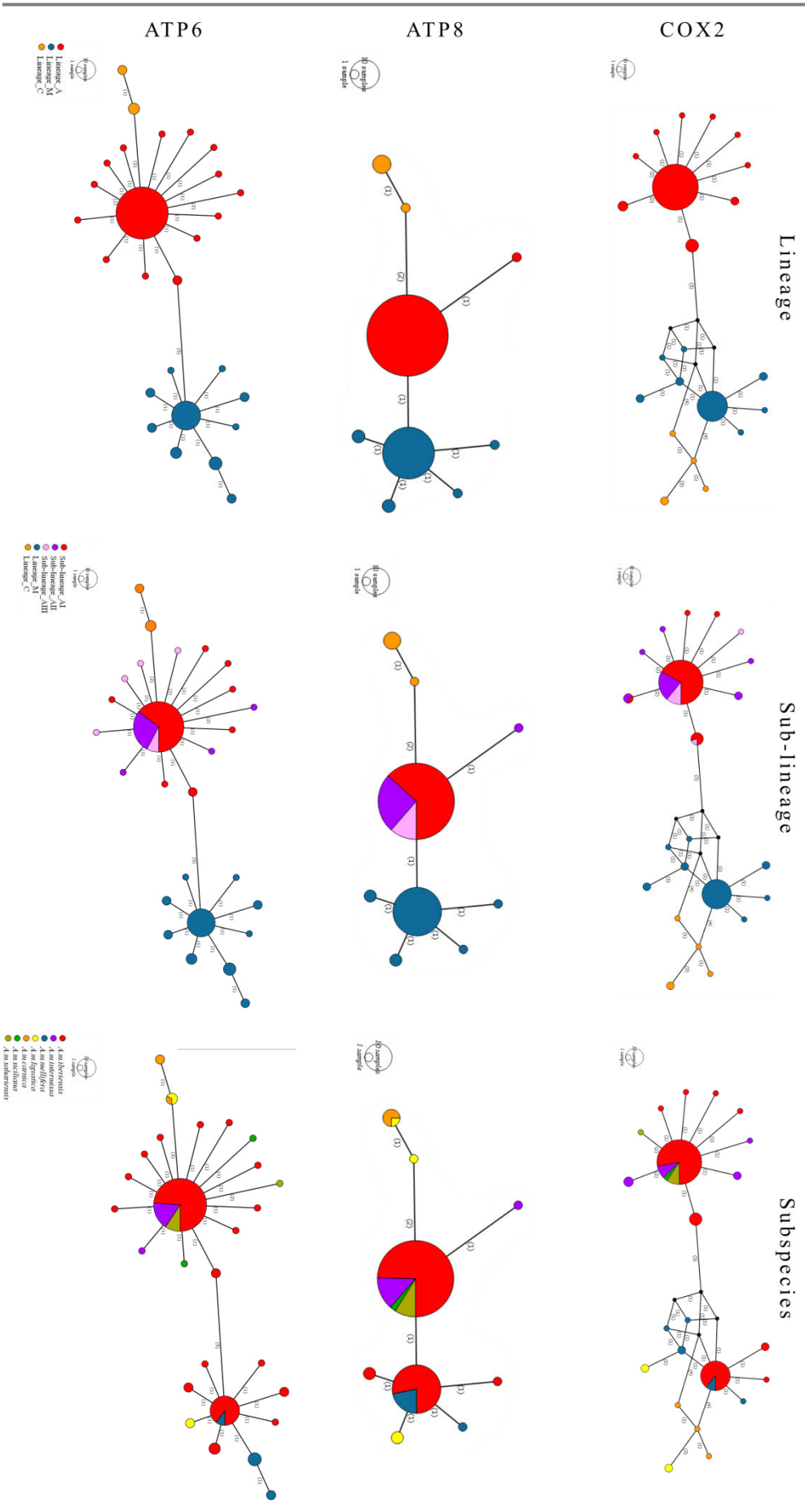
Table Sup VII-4-. Diversity measures, considering the single protein and ribosomal , the protein coding genes concatenated and the mitogenome. The individuals were divided by subspecies and for *A. m. iberiensis* and *A. m. ligustica* also by lineage. Ts., transitions; Tv., transversions; π, nucleotide diversity.

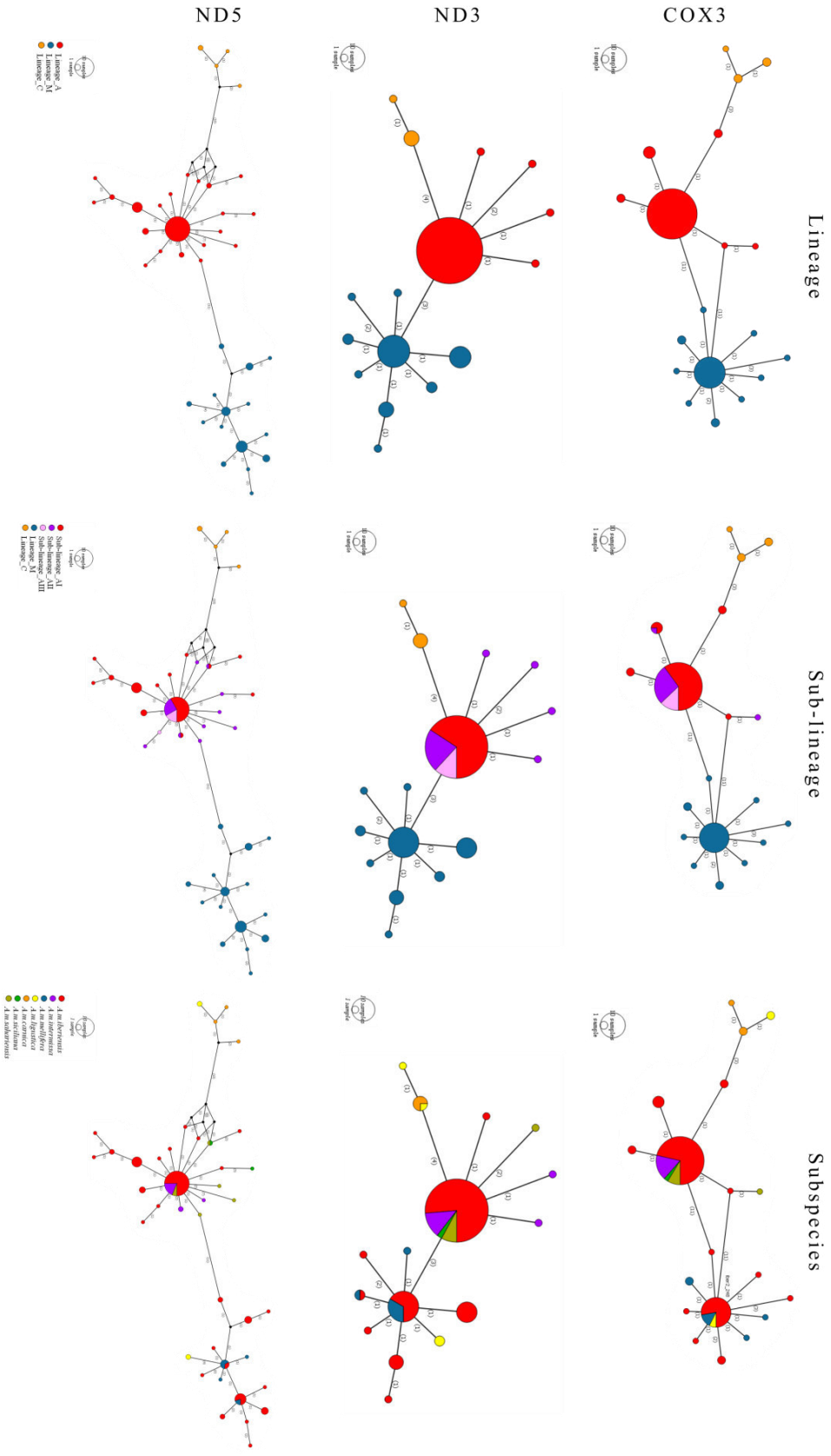
Table Sup VII-5- Genetic differentiation among the different groups. The individuals were divided by subspecies and for *A. m. iberiensis* and *A. m. ligustica* also by lineage. The diagonal elements, represents the average number of pairwise differences within population; above diagonal is the average number of pairwise differences between populations, while below diagonal corrected average pairwise difference. The values marked in bold have a p-value <0.05.

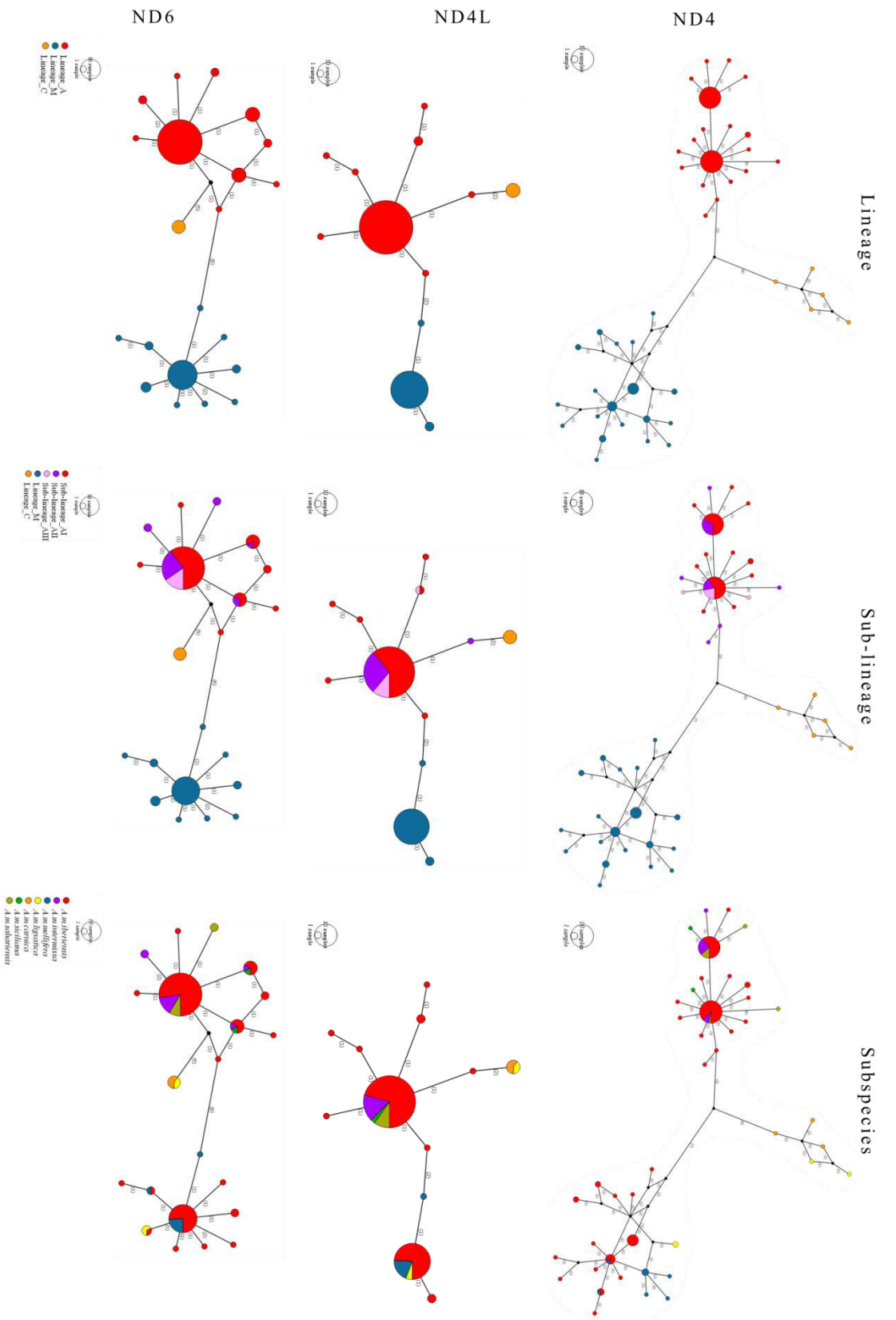
Table Sup VII-6- Similarities between the Bayesian phylogenetic trees measured by PH85 distance

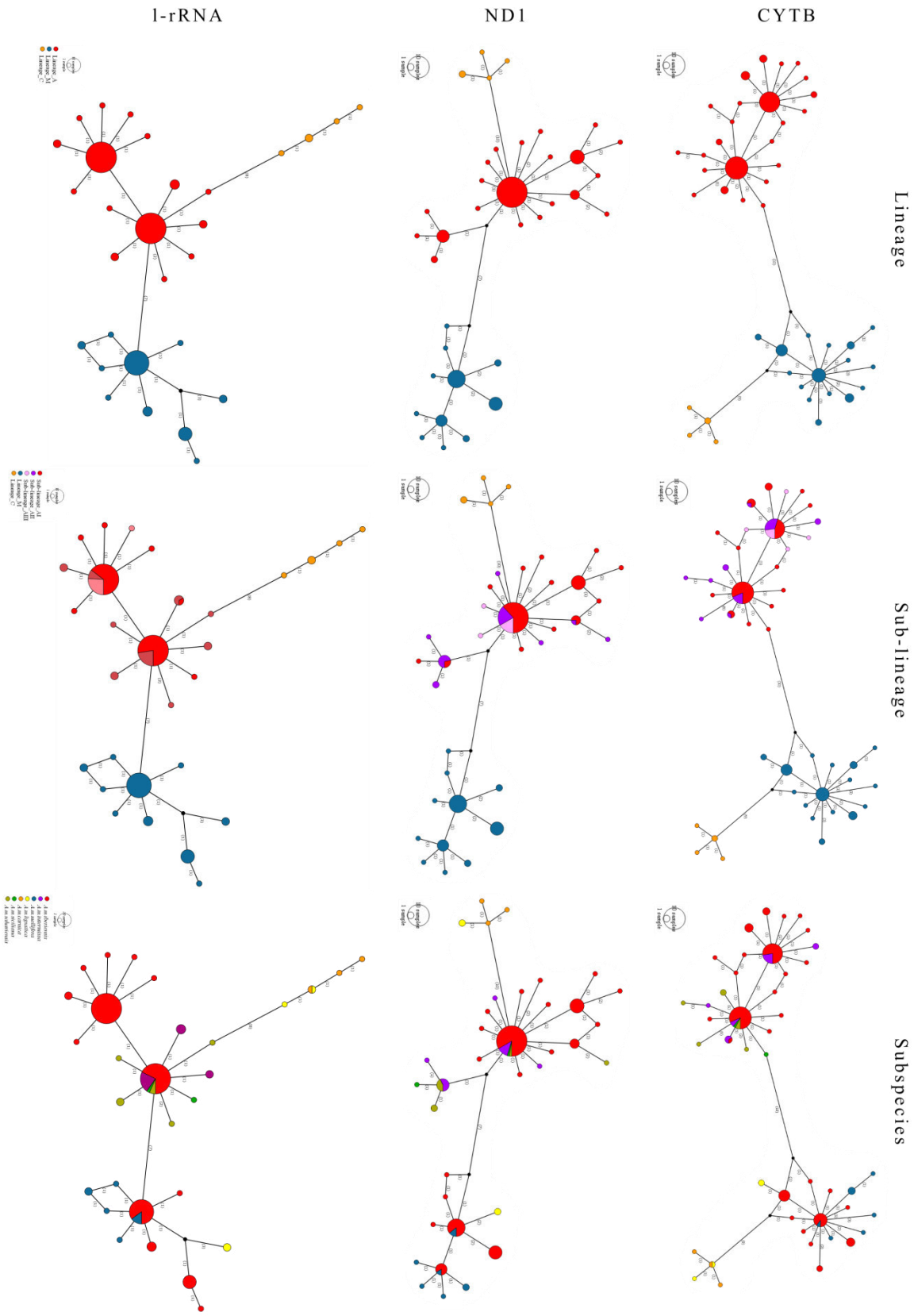
Table Sup VII-7- Number of groups identified for the mitogenome and single genes using ABGD with Kimura and Jukes-Cantor distance metrics and bPTP.


COX1
ND2
Mitogenome
Lineage
Sub-lineage
Subspecies


ATP6
ATP8
COX2
Lineage
Sub-lineage
Subspecies


ND5
ND3
COX3
Lineage
Sub-lineage
Subspecies


ND6
ND4L
ND4
Lineage
Sub-lineage
Subspecies
Lineage_A
Lineage_M
Lineage_C
Sub-lineage_AI
Sub-lineage_AII
Sub-lineage_AIII
Sub-lineage_M
Lineage_C
A.m.iberiensis
A.m.intermissa
A.m.iberica
A.m.carnica
A.m.siciliana
A.m.sahariensis


l-rRNA
ND1
CYTB
Lineage
Sub-lineage
Subspecies

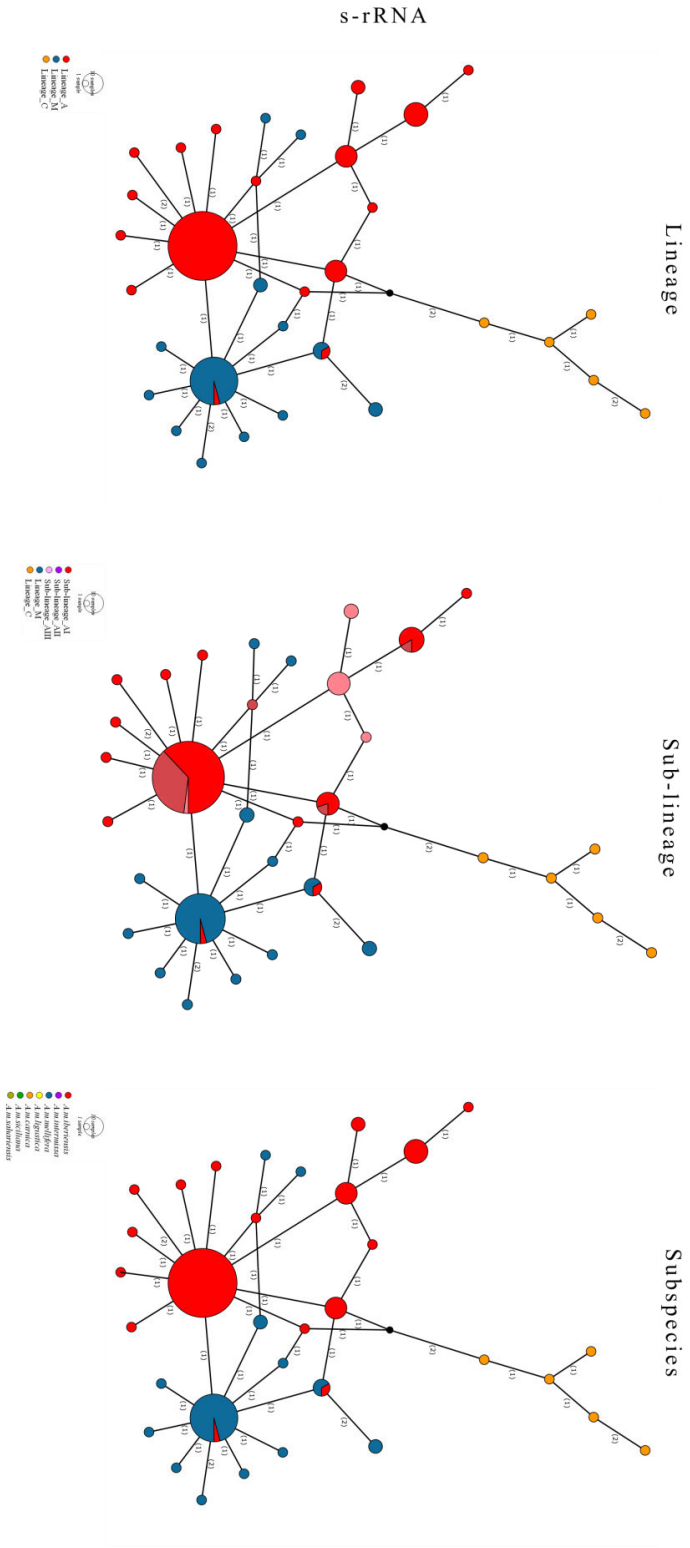


Figure VII-1. Median-joining networks inferred from different datasets. The size of circles is proportional to haplotype frequencies. Links between haplotypes are proportional to genetic distances between them. The colors of networks A) correspond to the three lineages; B) dividing the African lineage in sub-lineages, as identified by the *Dra*I test and C) to the different subspecies.

Published Papers

PLOS | ONE

RESEARCH ARTICLE

Reduced SNP Panels for Genetic Identification and Introgression Analysis in the Dark Honey Bee (*Apis mellifera mellifera*)

Irene Muñoz[1], Dora Henriques[1], J. Spencer Johnston[2], Julio Chávez-Galarza[1], Per Kryger[3], M. Alice Pinto[1]*

1 Mountain Research Centre (CIMO), Polytechnic Institute of Bragança, Campus de Sta. Apolónia, Apartado 1172, 5301–855, Bragança, Portugal, **2** Department of Entomology, Texas A&M University, College Station, Texas, 77843–2475, United States of America, **3** Aarhus University, Department of Agroecology, Forsøgsvej 1, 4200, Slagelse, Denmark

* apinto@ipb.pt

Citation: Muñoz I, Henriques D, Johnston JS, Chávez-Galarza J, Kryger P, Pinto MA (2015) Reduced SNP Panels for Genetic Identification and Introgression Analysis in the Dark Honey Bee (*Apis mellifera mellifera*). PLoS ONE 10(4): e0124365. doi:10.1371/journal.pone.0124365

Academic Editor: Wolfgang Blenau, University of Cologne, GERMANY

Received: January 9, 2015

Accepted: March 10, 2015

Published: April 13, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: Muñoz was supported by Fundación Séneca (Murcia, Spain) through the Post-doctoral fellowship 19149/PD/13-N whereas Dora Henriques and Julio Chávez-Galarza were supported by Fundação para a Ciência e Tecnologia (Portugal) through the PhD scholarships SFRH/BD/84195/2012 and SFRH/BD/68682/2010, respectively. This research was funded by Fundação para a Ciência e Tecnologia and COMPETE/QREN/EU through the projects PTDC/BIA-BEC/099640/2008 and

Abstract

Beekeeping activities, especially queen trading, have shaped the distribution of honey bee (*Apis mellifera*) subspecies in Europe, and have resulted in extensive introductions of two eastern European C-lineage subspecies (*A. m. ligustica* and *A. m. carnica*) into the native range of the M-lineage *A. m. mellifera* subspecies in Western Europe. As a consequence, replacement and gene flow between native and commercial populations have occurred at varying levels across western European populations. Genetic identification and introgression analysis using molecular markers is an important tool for management and conservation of honey bee subspecies. Previous studies have monitored introgression by using microsatellite, PCR-RFLP markers and most recently, high density assays using single nucleotide polymorphism (SNP) markers. While the latter are almost prohibitively expensive, the information gained to date can be exploited to create a reduced panel containing the most ancestry-informative markers (AIMs) for those purposes with very little loss of information. The objective of this study was to design reduced panels of AIMs to verify the origin of *A. m. mellifera* individuals and to provide accurate estimates of the level of C-lineage introgression into their genome. The discriminant power of the SNPs using a variety of metrics and approaches including the Weir & Cockerham's F_{ST}, an F_{ST}-based outlier test, Delta, informativeness (I_n), and PCA was evaluated. This study shows that reduced AIMs panels assign individuals to the correct origin and calculates the admixture level with a high degree of accuracy. These panels provide an essential tool in Europe for genetic stock identification and estimation of admixture levels which can assist management strategies and monitor honey bee conservation programs.

BiodivERsA/0002/2014. The open access publishing fees for this article have been covered by the Texas A&M University Online Access to Knowledge (OAK) Fund, supported by the University Libraries and the Office of the Vice President for Research. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

The role of introgression and admixture in conservation is a dilemma: While natural admixture may be an important evolutionary force in speciation and maintenance of genetic diversity [1–2], admixture induced by human activities may contribute, either directly or indirectly, to the extinction of many taxa [3]. Introduction of species, subspecies and habitat modifications has caused increased rates of admixture with native flora and fauna and introgression that can generate extinction and irretrievable loss of combinations of genotypes throughout the entire genome [4].

The honey bee, *Apis mellifera* L., represents a valuable model to study human-mediated change. Beekeeping has been practiced in Europe for many centuries [5], which has led to loss of native genetic diversity through three major mechanisms: (i) replacement of native populations by human-selected more docile and productive colonies, (ii) spread of honey bee pests and parasites, such as the mite *Varroa destructor* and the microsporidian *Nosema ceranae*, that have contributed to worldwide population declines [6–7], and (iii) recurrent introductions of commercial colonies (reviewed by De la Rúa et al. [8]).

The genetic diversity harbored in native honey bee subspecies is amongst the most important legacies that we can leave to future generations of beekeepers and farmers [9–10]. Native honey bee subspecies are important reservoirs of local adaptations; their extinction means the loss of unique combinations of traits shaped by natural selection over extended periods of time. These combinations can be important for a more sustainable beekeeping, as shown by a recent pan-European experiment [11].

In Europe, honey bees show considerable differences in morphological, behavioural and biological characters across their range as a result of historical patterns of isolation and adaptation to environmental conditions [8]. Those differences are materialized in 10 extant European subspecies, among the 30 subspecies currently recognized worldwide [12–16], representing thereby a substantial component of the total honey bee diversity. These 10 European subspecies have been grouped by morphological and molecular tools [12, 17–21] into two evolutionary lineages: the M-lineage, in Western Europe, and the C-lineage, in Eastern Europe.

Subspecies-specific genetic footprints can still be identified in Europe [22–28], in spite of centuries of beekeeping [5], although introgression and admixture events have also been detected in eastern [28–30] and western [9, 26, 31–32] European populations. The M-lineage *A. m. mellifera* (dark honey bee) has been recognized as the most threatened, with most of the threat due to introgression from the C-lineage [9, 31–32]. In addition to the documented intentional replacement of *A. m. mellifera* by *A. m. carnica* in Germany [33–34], the increasing trade of commercial breeds (mainly C-lineage *A. m. carnica*, *A. m. ligustica* and the hybrid buckfast) is threatening the genetic integrity of the native *A. m. mellifera* as many beekeepers prefer using commercial as opposed to native honey bees.

Increasing awareness that native honey bee diversity represents a valuable asset for sustainable beekeeping is fuelling local breeding and conservation efforts across Europe. One of the earliest, and until recently the single conservation program enacted by law, is that implemented by the Danish Beekeepers Association and the Læsø Beekeepers Association on behalf of the Danish Government in 1993 and the European Union in 1998 [35] to create a reserve and protect the *A. m. mellifera*. Following approval by the Scottish government of an order to protect the *A. m. mellifera* on the islands of Colonsay and Oronsay [The Bee Keeping (Colonsay and Oronsay) Order 2013], a second European reserve was recently created in the United Kingdom. Other *A. m. mellifera* conservation efforts, although not enacted by law, are underway in France, Holland, Norway, Switzerland, Ireland, and Belgium, among others (see the website "http://www.sicamm.org" run by the International Association for the Protection of the

European Dark bee). The success or failure of all these efforts will be tightly linked to efforts that monitor the integrity of these protected populations.

Assessing introgression is an important activity in honey bee breeding programs, especially when conservation of native subspecies is a major concern. This activity requires molecular tools that are reliable, inexpensive and preferably automated. Previous studies have monitored introgression between the endemic *A. m. mellifera* and introduced C-lineage subspecies using microsatellite and PCR-RFLP markers [31–32, 36]. However, with the publication of the honey bee genome [37], development of single-nucleotide polymorphism (SNP) markers [20, 38], and next generation sequencing becoming fast and affordable, particularly for a small genome as that of the honey bee (236 Mb), increasingly powerful tools are available to measure genomic ancestry and admixture levels occurring in both native and introduced honey bee populations [21, 39–40]. However, the genomic approach is not always cost-effective and low quality and/or degraded DNA can be a handicap to using genomic re-sequencing. Alternatively, ancestry can be estimated using a subset of highly informative SNPs ranging in number from a few dozens to several hundreds. The selected SNPs, commonly known as Ancestry-Informative Markers (AIMs), are those that exhibit large allele frequency differences between populations. AIMs can be used for inferring geographic origin of individuals [41–43], detecting illegal trade and translocation of animals [44], food authentication [45], for estimating overall admixture proportions efficiently and inexpensively [43, 46], among others. It is possible, using a panel of AIMs distributed throughout the genome, to estimate the relative ancestral proportions in admixed individuals, and infer the time since the admixture process [47–48].

The ability of an AIMs panel to measure ancestry is generally evaluated empirically, by examining its performance on a given set of samples for which ancestry is known [49]. In this paper, we employed five analytical methods to select different combinations of SNPs to form five nested panels of 48-, 96-, 144-, 192- and 384-AIMs optimized to estimate admixture proportions of C-lineage (*A. m. ligustica* and *A. m. carnica*) into the M-lineage *A. m. mellifera*. This was done in two successive stages. In the first stage, we evaluated the performance of the five selection methods [Weir & Cockerham's F_{ST}, an F_{ST}-based outlier test, Delta, informativeness (I_n), and PCA] on a training dataset, in an effort to select AIMs and to rank them by decreasing level of informativeness. In the second stage, we tested the power of the reduced five designed panels and validated their performance on holdout and simulated sets, by comparing the admixture estimates produced by the panels with those produced by an initial dataset of 1183 SNPs.

Material and Methods

Samples, DNA Extraction and SNP Genotyping

A total of 113 honey bee haploid males were collected in 2010 and 2011 across the native range of *A. m. mellifera*, *A. m. ligustica* and *A. m. carnica* in Europe (see the sampling map in Pinto et al. [9]). The samples of *A. m. mellifera* (N = 77) were collected from apiaries located in England (N = 8), France (N = 15), Belgium (N = 3), Denmark (N = 10), Holland (N = 15), Switzerland (N = 6), Scotland (N = 10), and Norway (N = 10) from protected and unprotected populations [9]. Colonies of protected populations have been identified by morphological (B. Dahle, pers. comm.) and molecular tools (mtDNA $tRNA^{leu}$-cox2 and microsatellites; [31, 32, 50–51]) as the best representatives of *A. m. mellifera* and have therefore been integrated into conservation programs. To prevent C-lineage introgression and assure pure breeding, these colonies have been maintained in islands or in isolated mating stations. Despite careful management to protect the threatened *A. m. mellifera* from C-lineage introgression, a recent SNP survey detected variable, although generally low, levels of introgression in these protected

populations (see Pinto et al. [9] for details). A reference collection of 36 samples representing C-lineage diversity was obtained from the natural range of *A*. *m*. *carnica* in Serbia (N = 8) and Croatia (N = 11) and from the natural range *A*. *m*. *ligustica* in Italy (N = 17). The owners of all the sampled apiaries gave permission to collect honey bee individuals from the hives. In each location, samples were taken from the inner part of hives, placed into absolute ethanol and stored at −20°C until molecular analysis.

Using a phenol/chloroform isoamyl alcohol (25:24:1) protocol [52], total DNA was extracted from the thorax of the 113 individuals, each representing a single colony. A total of 1536 SNP loci were genotyped for those individuals using Illumina's BeadArray Technology and the Illumina GoldenGate Assay with a custom Oligo Pool Assay (Illumina, San Diego, CA, USA) following manufacturer's protocols. The Oligo Pool consisted of the 1536 SNPs, which included the 768 most informative SNPs of Whitfield et al. [20] and 768 newly developed SNPs employed by Chávez-Galarza et al [38]. The 1536 SNP array was used previously to study diversity and introgression levels in populations of *A*. *m*. *mellifera* sampled across Western Europe [9] and to detect signatures of selection in the Iberian honey bee genome [38]. Genotype calling was performed using Illumina's GenomeStudio Data Analysis software. Of the initial 1536 SNPs, 353 did not meet the quality criteria for analysis and were therefore excluded from the dataset. The SNP filtering was as follows: 124 exhibited poorly separated intensity clusters or low signal intensity when visualized in the GenomeStudio software; 167 were monomorphic (defined by a cut-off criterion of >0.98 for the most common allele, as in Chávez-Galarza et al. [38]) across all populations; 54 did not map in the honey bee genome assembly Amel_4.0; and 8 hit two different genomic positions (the first with 100% identity and the second with 96–98%) in the honey bee genome assembly Amel_4.0 during the mapping process using the 100 bp flanking sequence. Allele frequencies were calculated for each of the remaining 1183 bi-allelic SNPs (S1 Table) in each population using the program Plink [53].

Selection of AIMs

Five different methods were employed on the initial 1183 SNP dataset for estimating marker information content. The first method, which has been one of the most popular for selecting informative loci, was the pairwise F_{ST} of Weir & Cockerham [54] as calculated at each locus using Genepop software [55]. The second method was the F_{ST}–based outlier test developed by Foll & Gaggiotti [56], which employs a Bayesian likelihood approach to detect loci deviating from neutral expectations (outliers). This outlier test was implemented in Bayescan 2.01 [56] using 20 pilot runs of 5 000 iterations (sample size of 5 000 and thinning interval of 10) and an additional burn-in of 50 000 iterations. The third method was based on the estimate of allele-frequency differential (Delta), which is one of the most straightforward ways to evaluate the information content of a SNP. For a bi-allelic marker, like a SNP, the Delta value is estimated as $|pA_i—pA_j|$, where pAi and pA_j are the frequencies of allele A in the i^{th} and j^{th} populations, respectively. When more than two populations were analyzed, the Delta value for each SNP locus was estimated as the mean across all pair-wise comparisons. The fourth method was the informativeness for assignment (I_n, natural logarithm of the number of populations) proposed by Rosenberg et al. [41]. I_n provides the amount of information gained about population assignment from observation of a single randomly chosen allele at a locus. This method assumes a uniform prior across K potential source populations for the origin of the allele. For a given set of populations, the minimum value of I_n (0) occurs when all alleles have equal frequencies in all populations whereas the maximum value (1) occurs when alleles are not shared among populations. I_n was calculated using the software Infocalc available at http://www.stanford.edu/group/rosenberglab/infocalc.html. Finally, the fifth selection method was principal component

analysis (PCA), which was performed using the PAST software [57]. The first eight principal components were used to calculate the information content of each SNP following the approach of Paschou et al. [58]. The loadings for each SNP were squared and summed over the eight most significant principal components to produce an estimate of informativeness.

SNPs were ranked and panels of SNPs tested using reference populations and the Anderson's Simple Training and Holdout method to reduce the potential for upward bias, which is introduced when loci are ranked and assessed using the same individuals [59]. To that end, a total of 34 pure (*sensu* Soland-Reckeweg et al. [32]) individuals of *A. m. mellifera*, previously identified in Pinto et al. [9], and all reference individuals (17 *A. m. ligustica* and 19 *A. m. carnica*) were used for SNP ranking (training set = 70) and the remaining 43 individuals of *A. m. mellifera* were reserved for panel testing (holdout set = 113). To minimize the effect of clusters of populations on the selection of the AIMs [41, 60–61], the five selection methods were tested using four training datasets. The first dataset consisted of 70 individuals: 34 pure *A. m. mellifera* and 36 C-lineage individuals, with no distinction between the *A. m. carnica* and *A. m. ligustica* subspecies (dataset I). The second dataset consisted of 51 individuals: 34 pure *A. m. mellifera* and 17 *A. m. ligustica* (dataset II). The third dataset consisted of 53 individuals: 34 pure *A. m. mellifera* and 19 *A. m. carnica* (dataset III). Finally, the fourth dataset consisted of 70 individuals: 34 pure *A. m. mellifera*, 17 *A. m. ligustica* and 19 of *A. m. carnica* (dataset IV).

Ranking of SNPs

The five selection methods were implemented on the four training datasets producing a total of 20 information content values for each of the 1183 SNPs. These values were ranked and analyzed individually and then were averaged in two steps to obtain a single global value per SNP. In the first step the information content values were averaged across the four training datasets for each of the five selection methods. In the second step the information content values produced by each selection method were converted into a 0–1 scale and then averaged to obtain a global score for each of the 1183 SNPs. After standardizing the values produced by the five selection methods, the global ranking was obtained for the 1183 SNPs using the global score. Given that linked loci yield redundant information, having therefore similar resolving power, markers were excluded if they were within a predefined genetic distance (<1 cM) of higher ranking selected SNPs. The genetic distance of the remaining SNPs ranged from 1.01 to 24.25 cM with a mean of 4.64 cM. Prior to obtaining the global score for each SNP, pairwise associations between information content values produced by the five methods and between the four training datasets were calculated using the Spearman`s rank correlation coefficient, in order to compare the five selection methods and examine the effect of clusters of populations.

Panel Testing

Five panels of 48-, 96-, 144-, 192- and 384-SNPs (sets defined by multiplex sizes of commercial assays) were designed from the top-ranked SNPs. These nested panels were tested against a holdout set and a simulated set to obtain the admixture proportions estimated by each SNP panel. The holdout set (113 individuals) consisted of 34 pure individuals plus 43 reserved individuals of *A. m. mellifera* and the reference *A. m. ligustica* (17 individuals) and *A. m. carnica* (19 individuals), as described above. The simulated set (1000 individuals) was generated with the program ONCOR [62] using the function "simulate a single mixture". Ten populations, each with 100 simulated genotypes, were simulated using different levels of introgression (0, 1, 5, 10, 20, 30, 40, 50, 75, and 90%).

Two approaches were used to validate the five reduced AIMs panels. First, a PCA was performed with SNPs in each AIMs panel on the holdout set using the software PAST to generate

two-dimensional PCA and to visualize the stability of population assignment produced by the panels. Second, ancestry and admixture was analyzed. Admixture proportions were estimated with SNPs in each AIMs panel for the holdout and simulated sets using a model-based maximum likelihood estimation of individual ancestries implemented in the software Admixture v1.23 [63]. Coancestry spanning 1–6 populations (K = 1–6, using the default termination criterion that stops the runs when the log-likelihood increases by less than $\varepsilon < 0.0001$ between iterations) was explored for each AIMs panel and the optimal K was identified with the inferred number of populations producing the lowest cross-validation error (CV) during the clustering analysis.

The performance of each reduced panel was examined using different approaches. First, the pairwise differences between admixture proportions inferred from the initial 1183 SNP dataset and the five panels were tested using a Mann-Whitney test. Second, the precision of each panel was tested against the initial 1183 SNP dataset by calculating linear regression coefficients (r^2) and the standard deviations of the differences between admixture proportions. Finally, the accuracy of the reduced panels was estimated via percentage of absolute error of admixture estimates obtained with the five panels in relation to the initial 1183 SNP dataset.

Results

Identification and Ranking of AIMs

The majority of the 1183 SNPs assessed in this study using five selection methods (pairwise Weir & Cockerham's F_{ST}, F_{ST}-based outlier test, Delta, I_n and PCA) contain high levels of information content (Fig 1, S2 Table), facilitating the design of reduced panels for genetic identification and introgression analysis in the dark honey bee, *A*. *m*. *mellifera*. The distribution of frequency histograms and percentiles of genetic information content of the 1183 SNPs estimated by each selection method and training dataset are shown in Fig 1. The 50th percentile ranges of the four training datasets were 0.6974–0.7712, 0.5459–0.6362, 0.5532–0.7601, 0.3345–0.3583 and 0.0038–0.0040 for the Weir & Cockerham's F_{ST}, F_{ST}-based outlier test, Delta, I_n and PCA, respectively, indicating a high level of information content for most SNPs and a similar pattern among the four training datasets (Fig 1).

The level of similarity (Spearman's rank correlation, r_s) between the different estimates of genetic information content produced by the five selection methods across the four training datasets is shown in Table 1. The highest correlation values were observed for Weir & Cockerham's F_{ST}, Delta and I_n ($0.7648 \leq r_s \leq 0.9985$, $P<0.001$) whereas a moderate correlation was detected between the F_{ST}-based outlier test and Weir & Cockerham's F_{ST}, Delta and I_n ($0.2864 \leq r_s \leq 0.6592$, $P<0.001$). The lowest correlations were observed between PCA and the other four methods ($-0.2228 \leq r_s \leq 0.1025$, $0.000 \leq P \leq 0.9412$). Regarding the four training datasets (Fig 2), high correlation values were observed across selection methods ($0.7557 \leq r_s \leq 0.9727$, $P<0.001$).

Using an information content cutoff value ≥ 0.25, which indicates very great genetic differentiation [64], a total of 627 AIMs were identified by the methods of Weir & Cockerham's F_{ST}, Delta, I_n, and F_{ST}-based outlier test. Of these, the top-ranked 384 AIMs were selected using the five methods and the four training datasets. The extent of overlap of the 384 AIMs across the five selection methods and the four training datasets is shown in Fig 2. Overlap between any two methods and across datasets ranged between 382 (Weir & Cockerham's F_{ST} and Delta for dataset I; Fig 2A) and 134 (Delta and PCA for dataset III; Fig 2C). The number of AIMs that were simultaneously selected by the five methods was lower, ranging from 82 (dataset I; Fig 2A) to 97 (dataset IV; Fig 2D). A substantially higher amount of overlap (273 AIMs; Fig 2E), supported by high correlation values ($r_s \geq 0.7557$, $P<0.001$; Fig 2F), was observed across the

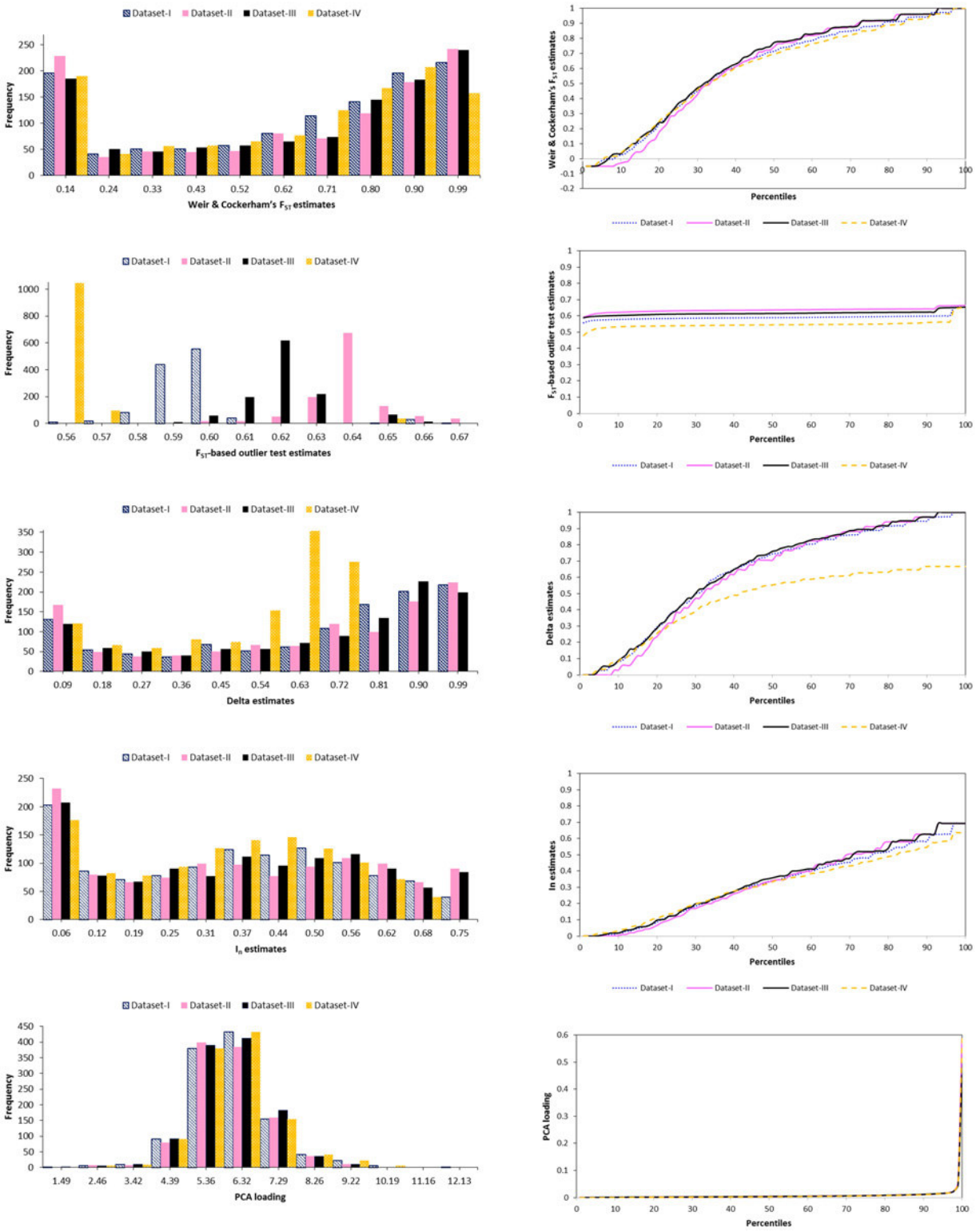


Fig 1. Frequency histograms and percentiles of the estimates of genetic information contained in the initial 1183 SNP dataset. Information content produced by the five selection methods (pairwise Weir & Cockerham's F_{ST}, F_{ST}-based outlier test, Delta, I_n and PCA) is shown for the four training datasets (I, II, III and IV).

doi:10.1371/journal.pone.0124365.g001

Table 1. Comparison of selection methods and training datasets.

		Dataset I: *A. m. mellifera* & C lineage (*A. m. ligustica* & *A. m. carnica* together)					Dataset II: *A. m. mellifera* & *A. m. ligustica*					Dataset III: *A. m. mellifera* & *A. m. carnica*					Dataset IV: *A. m. mellifera A. m. ligustica* & *A. m. carnica*				
		F_{ST}	Delta	I_n	PCA	F_{ST} outlier test	F_{ST}	Delta	I_n	PCA	F_{ST} outlier test	F_{ST}	Delta	I_n	PCA	F_{ST} outlier test	F_{ST}	Delta	I_n	PCA	F_{ST} outlier test
Dataset I	F_{ST}		0.0000	0.0000	0.7134	0.0000	0.0000	0.0000	0.0000	0.0620	0.0000	0.0000	0.0000	0.0000	0.6130	0.0000	0.0000	0.0000	0.0000	0.7134	0.0000
	Delta	0.9985		0.0000	0.6139	0.0000	0.0000	0.0000	0.0000	0.0439	0.0000	0.0000	0.0000	0.0000	0.6772	0.0000	0.0000	0.0000	0.0000	0.6139	0.0000
	I_n	0.9977	0.9957		0.7284	0.0000	0.0000	0.0000	0.0000	0.0753	0.0000	0.0000	0.0000	0.0000	0.4980	0.0000	0.0000	0.0000	0.0000	0.7284	0.0000
	PCA	0.0107	0.0147	0.0101		0.0000	0.9412	0.8104	0.7135	0.0000	0.0000	0.0390	0.0241	0.0357	0.0000	0.0711	0.4901	0.1065	0.5267	0.0000	0.0001
	F_{ST} outlier test	0.5974	0.5778	0.6280	-0.1195		0.0000	0.0000	0.0000	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Dataset II	F_{ST}	0.9364	0.9370	0.9392	-0.0021	0.5666		0.0000	0.0000	0.3782	0.0000	0.0000	0.0000	0.0000	0.8227	0.0000	0.0000	0.0000	0.0000	0.9412	0.0000
	Delta	0.9338	0.9357	0.9313	-0.0070	0.5205	0.9906		0.0000	0.3980	0.0000	0.0000	0.0000	0.0000	0.7760	0.0000	0.0000	0.0000	0.0000	0.8104	0.0000
	I_n	0.9329	0.9324	0.9342	-0.0107	0.5581	0.9923	0.9960		0.5187	0.0000	0.0000	0.0000	0.0000	0.8981	0.0000	0.0000	0.0000	0.0000	0.7135	0.0000
	PCA	0.0543	0.0586	0.0517	0.8545	-0.1074	0.0256	0.0246	0.0188		0.0000	0.0017	0.0004	0.0008	0.0000	0.1510	0.0986	0.0426	0.1329	0.0000	0.0000
	F_{ST} outlier test	0.4404	0.4235	0.4637	-0.2018	0.8128	0.4998	0.4732	0.5124	-0.2228		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Dataset III	F_{ST}	0.9328	0.9317	0.9353	0.0600	0.5751	0.7965	0.7705	0.7719	0.0909	0.3301		0.0000	0.0000	0.2647	0.0000	0.0000	0.0000	0.0000	0.0390	0.0000
	Delta	0.9347	0.9353	0.9327	0.0656	0.5337	0.7825	0.7690	0.7663	0.1025	0.2864	0.9925		0.0000	0.4242	0.0000	0.0000	0.0000	0.0000	0.0241	0.0000
	I_n	0.9331	0.9312	0.9349	0.0611	0.5742	0.7829	0.7648	0.7656	0.0969	0.3150	0.9943	0.9960		0.3212	0.0000	0.0000	0.0000	0.0000	0.0357	0.0000
	PCA	-0.0147	-0.0121	-0.0197	0.6450	-0.1333	-0.0065	0.0083	0.0037	0.5534	-0.1435	-0.0325	-0.0233	-0.0289		0.0000	0.3690	0.3671	0.4471	0.0000	0.0000
	F_{ST} outlier test	0.5524	0.5361	0.5793	-0.0525	0.8862	0.4451	0.3910	0.4189	-0.0418	0.6063	0.6337	0.6006	0.6444	-0.1440		0.0000	0.0000	0.0000	0.0711	0.0000
Dataset IV	F_{ST}	0.9883	0.9862	0.9875	0.0201	0.6028	0.9301	0.9216	0.9222	0.0480	0.4567	0.9411	0.9387	0.9384	-0.0261	0.5771		0.0000	0.0000	0.4901	0.0000
	Delta	0.9459	0.9485	0.9474	0.0470	0.5487	0.9177	0.9022	0.9014	0.0590	0.4237	0.9241	0.9169	0.9152	-0.0262	0.5475	0.9683		0.0000	0.1065	0.0000
	I_n	0.9825	0.9798	0.9832	0.0184	0.6081	0.9313	0.9277	0.9308	0.0437	0.4733	0.9273	0.9279	0.9300	-0.0221	0.5798	0.9948	0.9710		0.5267	0.0000
	PCA	0.0107	0.0147	0.0101	1.0000	-0.1195	-0.0021	-0.0070	-0.0107	0.8545	-0.2018	0.0600	0.0656	0.0611	0.6450	-0.0525	0.0201	0.0470	0.0184		0.0001
	F_{ST} outlier test	0.6043	0.5867	0.6267	-0.1154	0.9119	0.5549	0.5292	0.5634	-0.1250	0.8250	0.5819	0.5589	0.5943	-0.1376	0.8758	0.6344	0.6104	0.6592	-0.1154	

Spearman's rank correlation coefficients (lower triangle), and corresponding *P*-values (upper triangle), between information content estimates produced by the five selection methods (Weir & Cockerham's F_{ST}, Delta, informativeness (I_n), PCA, F_{ST}-based outlier test) using the four training datasets (I, II, III and IV).

doi:10.1371/journal.pone.0124365.t001

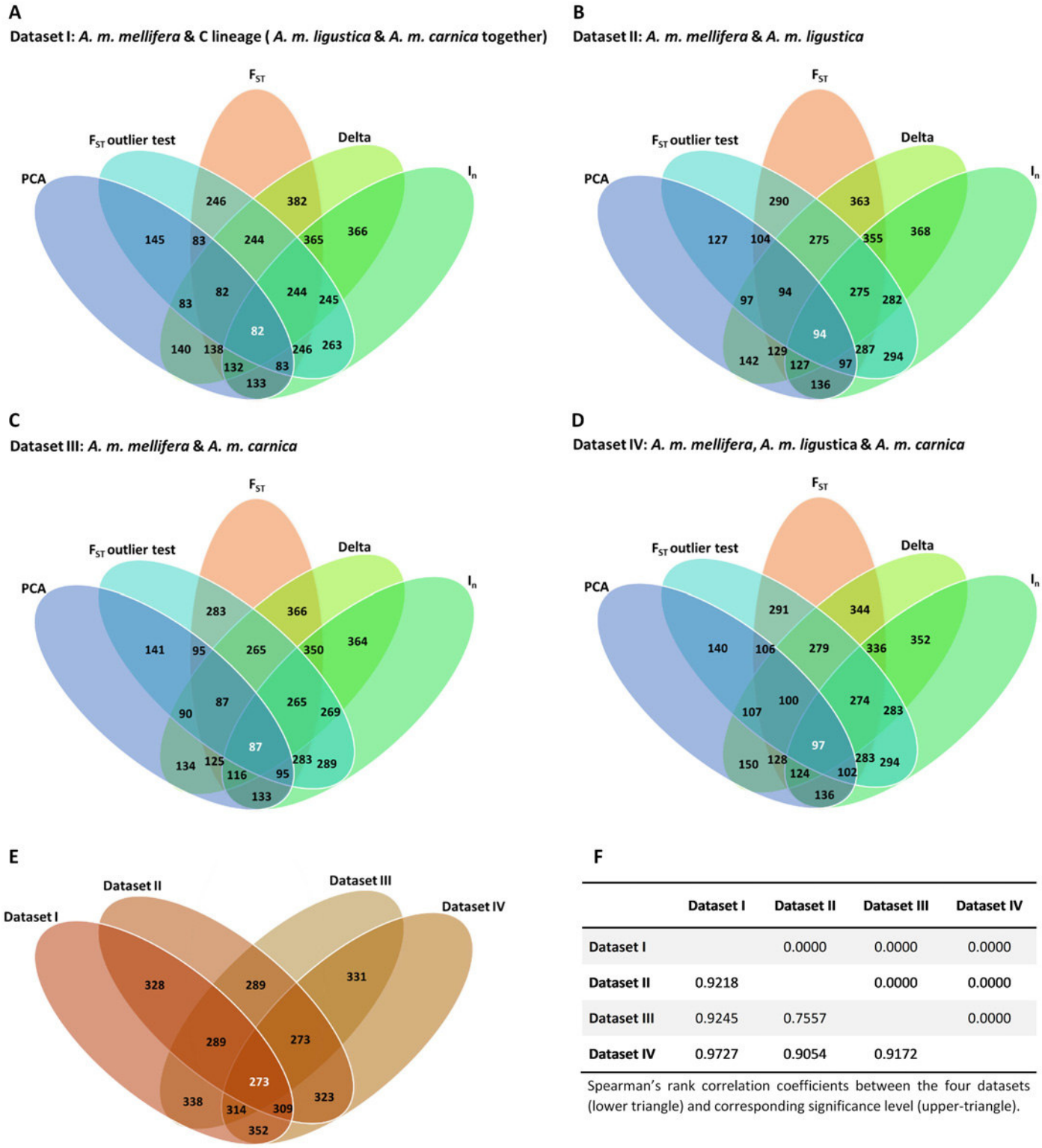


Fig 2. (A-E) Venn diagrams showing the extent of overlap of the top-ranked 384 AIMs. (A-D) Overlap among the five selection methods (pairwise Weir & Cockerham's F_{ST}, F_{ST}-based outlier test, Delta, I_n and PCA) and the four training datasets (I, II, III and IV). (E) Overlap among the four training datasets, after averaging the information content obtained with the five selection methods, and (F) corresponding Spearman's rank correlation coefficients.

doi:10.1371/journal.pone.0124365.g002

four training datasets, suggesting that the different population groupings have a small effect on the AIMs ranking. The global ranking of the 384 AIMs was used to design reduced panels of 192-, 144-, 96-, and 48 that included SNPs with the highest respective global scores. The performance of these reduced panels was subsequently assessed using the holdout and simulated sets.

Validation of the AIMs Panels

The performance of the five AIMs panels (48-, 96-, 144-, 192-, 384-AIMs) was first validated by using PCA to produce a visual summary of the observed genetic variation carried by the holdout set (Fig 3). The overall diversity pattern is characterized by the presence of two distinct

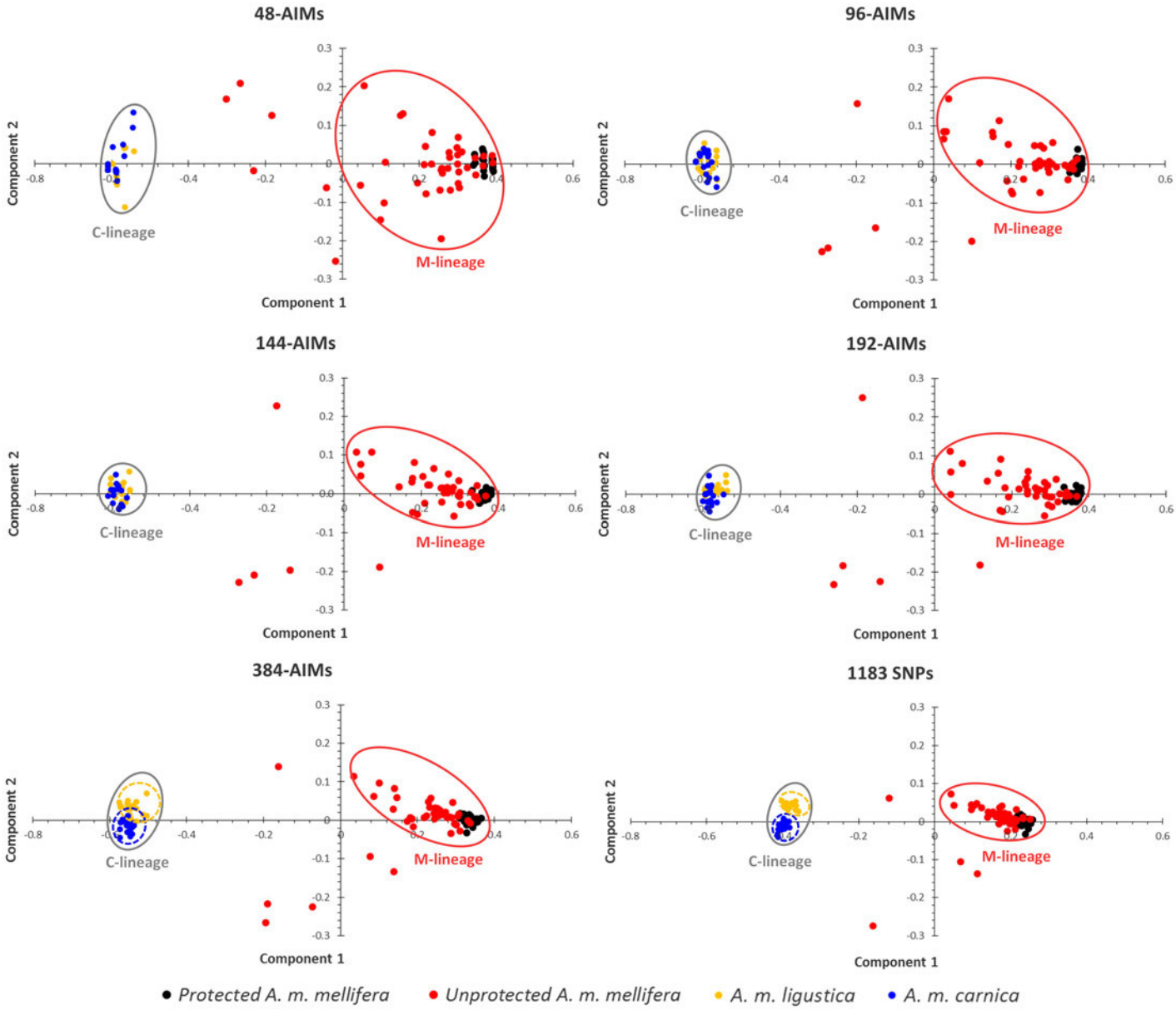


Fig 3. Principal components analysis. Plots obtained for the holdout set using the five AIMs panels (48-, 96-, 144-, 192-, 384-AIMs) and the initial 1183 SNP dataset.

doi:10.1371/journal.pone.0124365.g003

clusters, which are coincidental with the M and C evolutionary lineages. This pattern was captured by every single AIMs panel, although a greater dispersion was observed for the smaller panels. Additionally, the panels with less than 192 AIMs were unable to distinguish the two C-lineage subspecies, *A.m. ligustica* and *A.m. carnica*, which were clearly identified by the initial 1183 SNPs and, to a lesser degree, by the 384-AIMs panel.

Ancestry and admixture analyses based on admixture estimates confirm the overall pattern captured by the PCA (S3 Table and S1 Fig). At the optimal K = 2 (inferred by the initial 1183 SNP dataset and the five AIMs panel), the two clusters corresponded to the C and M-lineages. However, C-lineage individuals formed a more homogeneous cluster than those of the M-lineage individuals. While membership proportions in the C-lineage cluster were greater than 95% for the five AIMs panels, the M-lineage cluster comprised 13 (384-AIMs and 1183 SNPs), 14 (48- and 192-AIMs) and 15 (96- and 144-AIMs) individuals with membership proportions lower than 85%, a pattern that was already evident in the PCA plots.

The introgression levels exhibited by individuals of the M-lineage cluster were significantly higher (Student's t-test, $P<0.001$) in unprotected (13.76–15.18%, with 1183 SNPs and 48 AIMs, respectively) than in protected individuals (0.08–0.52%, with 96 AIMs and 1183 SNPs, respectively) for any AIMs panel. The overall estimates of C-lineage introgression into *A. m. mellifera* varied with the panel (8.4, 7.9, 7.8, 7.9, 7.5 and 7.7% with 48-, 96-, 144-, 192-, 384-AIMs and 1183 SNPs, respectively), although the differences were not statistically significant (Mann-Whitney test, $0.8225 \leq P \leq 0.9983$; S4 Table).

In addition to the admixture analyses using the holdout set, the AIMs panels were further validated using a simulated set of 10 different levels of C-lineage introgression (0, 1, 5, 10, 20, 30, 40, 50, 75, and 90%). As for the analyses with the holdout set, the simulated set produced two clusters corresponding to M and C lineages with no significant differences in admixture proportions between the different AIMs panels and the initial 1183 SNP dataset (Mann-Whitney test, $P \geq 0.2313$; S5 Table).

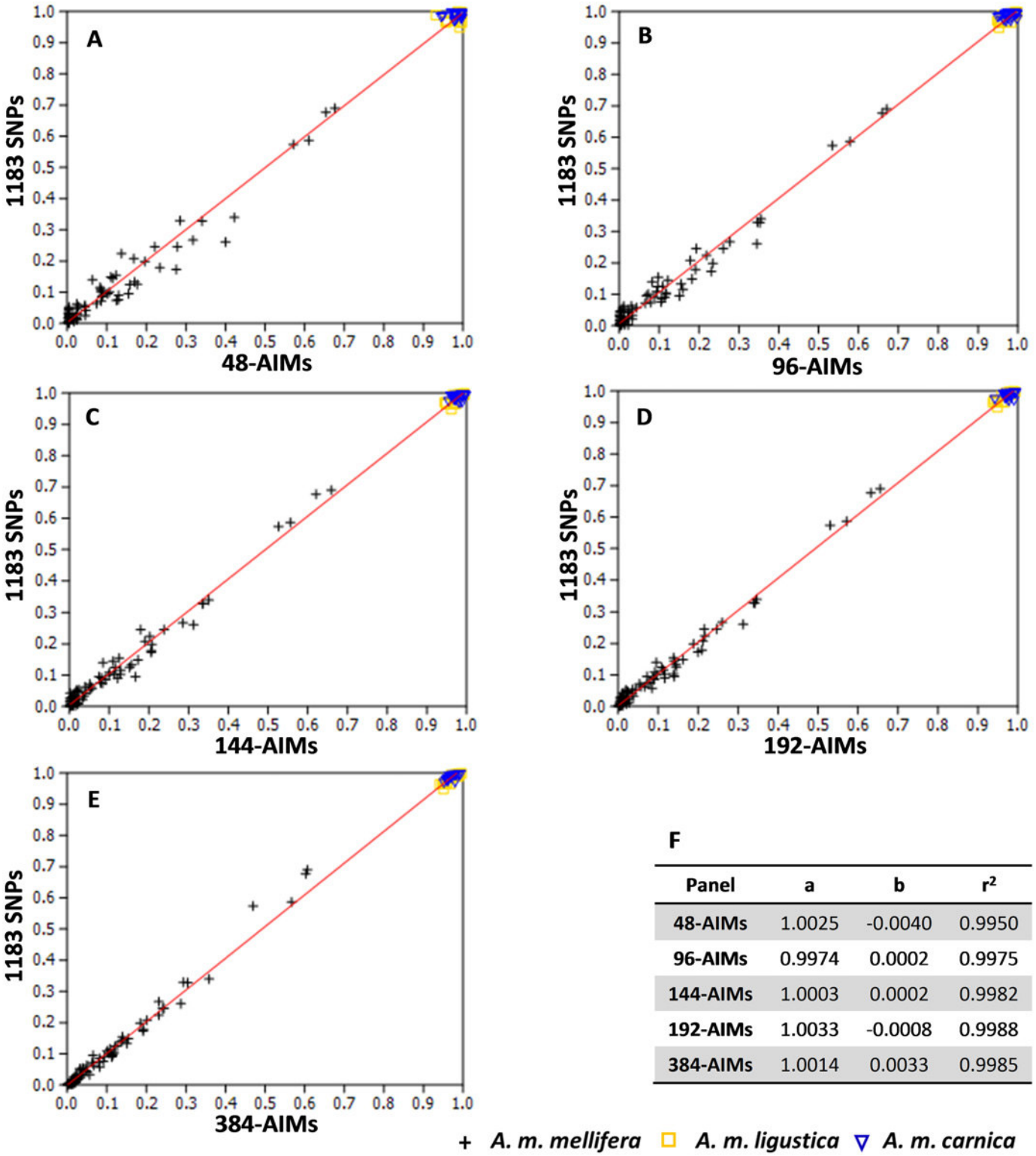
Assignment's precision and accuracy

The power of the reduced AIMs panels in identifying *A. m. mellifera* and estimating admixture proportions was evaluated on the holdout set. Estimates of C-lineage introgression into *A. m. mellifera* inferred from the five panels were greatly concordant with those inferred from the initial 1183 SNP dataset, as indicated by the high correlation values ($r \geq 0.997$; Fig 4). Despite the high correlations obtained for each comparison, the error rate in admixture estimates, which is very low for all the panels (0.0012–0.0042 with the simulated set and 0.4–1.3 with the holdout set), does increase as the size of the panel decreases (S2 Fig). Nevertheless, the reduced AIMs panels provide good precision in estimating admixture proportions.

As another assessment of the performance of the panels, the accuracy was calculated via absolute error. The success of assignment of the 113 individual genotypes of the holdout set to genetic origin and level of admixture inferred from the different AIMs panels is shown in Fig 5. The average percentage of correct assignment was high varying from 98.2, 98.8, 99.0, 99.2 to 99.4% for the 48-, 96-, 144-, 192- and 384-AIMs panels, respectively. The chosen AIMs panels accurately distinguish M/C admixture, therefore these results suggest that a small number of AIMs are sufficient to identify *A. m. mellifera* and estimate introgression from C-lineage colonies with great accuracy.

Discussion

The recognition that native honey bee genetic diversity is fundamental for sustainable beekeeping and for facing the challenges of a rapidly changing world (e.g. climate change, novel



F

Panel	a	b	r^2
48-AIMs	1.0025	-0.0040	0.9950
96-AIMs	0.9974	0.0002	0.9975
144-AIMs	1.0003	0.0002	0.9982
192-AIMs	1.0033	-0.0008	0.9988
384-AIMs	1.0014	0.0033	0.9985

Fig 4. Linear regression. (A-E) Plots between admixture proportions inferred from the initial 1183 SNP dataset and those inferred from the five AIMs panels (48-, 96-, 144-, 192-, 384-AIMs) using individuals of the holdout set. (F) Parameters and coefficients for each AIMs panel.

doi:10.1371/journal.pone.0124365.g004

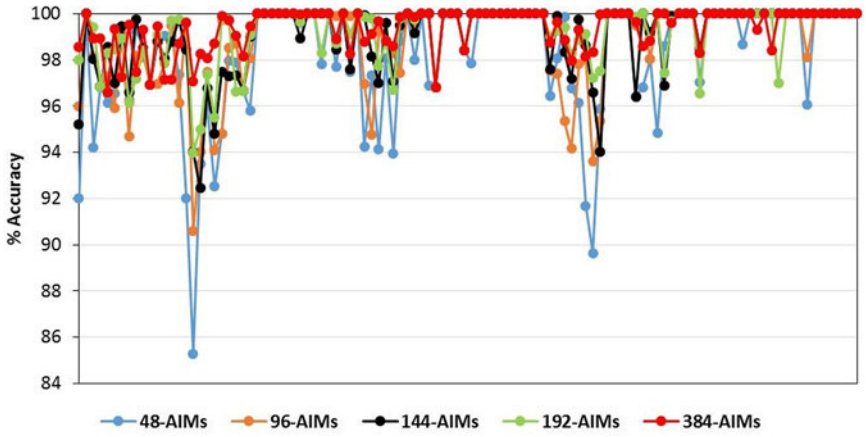


Fig 5. Assignment accuracy. Percentage obtained with the five AIMs panels (48-, 96-, 144-, 192-, 384-AIMs) for each of the 113 individuals of the holdout set.

doi:10.1371/journal.pone.0124365.g005

diseases and parasites) is stimulating implementation of conservation programs across Europe in an attempt to recover and protect *A. m. mellifera*, which is the European honey bee subspecies with the widest natural range [12], and at the same time the most threatened by introgression [9, 31]. The need of a reliable, high-throughput, and cost-effective tool for identifying candidate *A. m. mellifera* colonies targeted for conservation, a crucial step when managing conservatoires, motivated the design of reduced AIMs panels containing the most informative SNPs to verify ancestry and introgression from C-lineage subspecies. In this study we developed, validated and tested the first reduced AIMs panels for honey bees. Our results provide strong confidence in a panel of 384 AIMs and show that even smaller subsets of 192-, 144-, 96- and 48-AIMs are able to identify ancestry and estimate introgression with great accuracy. These reduced panels promise to be a useful tool for routine identification of *A. m. mellifera* colonies maintained in the breeding populations of conservation programs.

The AIMs included in the five reduced panels were simultaneously selected by pairwise Weir & Cockerham's F_{ST}, F_{ST}-based outlier test, Delta, I_n and PCA, in order to balance out the limitations of each individual method [41, 58, 65]. These selection methods have proved to be powerful, although with varying performances, in identifying population informative markers in a wide range of organisms [43, 58, 60–61, 65]. A great extent of overlap of top-ranked AIMs was obtained for the five selection methods, especially for pairwise Weir & Cockerham's F_{ST}, Delta, and I_n suggesting that they capture the same information. Nonetheless, the smaller panels (48-, 96-, 144-, 192-AIMs) did not necessarily include all AIMs simultaneously detected by the five methods as the global ranking depended on the average score. High pairwise correlation values were obtained for Weir & Cockerham's F_{ST}, Delta and I_n but not for PCA, as found by Wilkinson et al. [65]. PCA has been recommended for ranking markers because it has the advantage of generating an overall estimate for a single SNP locus whereas the other methods require estimate of an average from pairwise calculations when the number of populations is greater than two [58].

The five reduced panels tested with the holdout and simulated sets performed virtually as well as the initial 1183 SNP dataset, as revealed by the strong correlations obtained between admixture estimates and low associated error rates. The assignment power was high across the five panels with average values of correct assignment varying between 98.2 and 99.4%, although the accuracy decreased slightly with panel size. Nonetheless, even the 48-AIMs panel exhibited high accuracy levels, which is not surprising as it includes the AIMs with the greatest resolution

power. Studies on other organisms have also found good performances with panels of similar sizes [43, 45, 60, 65], detecting sharp drops in accuracy for a number of SNPs below 25 [45, 60].

Evaluation of different combinations of the focal *A. m. mellifera* and the two most common sources of foreign genes, *A. m. ligustica* and *A. m. carnica*, revealed a negligible effect of population groupings on the AIMs ranking. These results suggest that the designed panels are suited for identifying and assessing introgression of *A. m. ligustica*, *A. m. carnica* or both into *A. m. mellifera*. While these panels will possibly perform well in the presence of other C-lineage subspecies, more complex combinations that include sources of different evolutionary lineages will require further testing and, most likely, new panels developed from broader baseline datasets. Additionally, it should be noted, that these reduced panels are not suitable for standard population genetic analyses, including determining allelic diversity or measuring isolation by distance, genetic drift or bottleneck effect. The bias introduced through selection for markers that segregate among target populations would seriously compromise these calculations [66–67].

Ancestry identification of honey bee subspecies is undergoing steady development (reviewed by Meixner et al. [68]) from classical morphometry, analysis of allozymes, mitochondrial DNA, nuclear microsatellites, and now SNP tools. Because researchers must balance the cost of genotyping many samples versus many loci, herein we developed five nested reduced panels that include AIMs with the highest resolution power for discriminating subspecies of the divergent M and C evolutionary lineages. While the 384-AIMs panel is also capable of discriminating the C-lineage *A. m. ligustica* and *A. m. carnica*, for estimating C-lineage introgression into *A. m. mellifera* we recommend using the 96-AIMs panel because it is accurate; and high-throughput 96-plex genotyping assays can be outsourced at an affordable cost ($8 900 for 480 samples), representing a saving of 92.4% when compared with the 1536-plex assay ($116 800 for 480 samples).

In conclusion, the proposed AIMs panels can be actively used as a tool in conservation management of *A. m. mellifera* populations that suffer from hybridization and introgression with the most commonly introduced and beekeepers' preferred *A. m. ligustica* and *A. m. carnica* subspecies. This can be an important advance because the current European regulation on organic beekeeping states that "preference shall be given to the use of European breeds of *Apis mellifera* and their local ecotypes" and several conservation programs have been undertaken in Europe (reviewed by De la Rúa et al. [8]). The use of these panels will apply well to monitoring, management and conservation programs of *A. m. mellifera* in Western Europe, which usually require high-sample throughput, and will be a resource for the honey bee community to obtain accurate genetic information at reduced costs.

Supporting Information

S1 Fig. Ancestry estimates. Global estimates (y-axis), for the 113 individuals of the holdout set (x-axis), inferred from the five AIMs panels (48-, 96-, 144-, 192-, 384-AIMs) and the initial 1183 SNP dataset using the model-based approach implemented in the ADMIXTURE software. Results are shown for the optimal $K = 2$, which distinguishes the M (red) and C (cyan) evolutionary lineages of *A. mellifera*.
(TIFF)

S2 Fig. Standard deviation (SD) of admixture proportions. Precision estimates obtained using the SD of the differences between admixture proportions inferred from the initial 1183 SNP dataset and the five AIMs panels (48-, 96-, 144-, 192-, 384-AIMs) using the holdout (blue line) and simulated (orange line) sets.
(TIFF)

S1 Table. Input file containing the 1183 coded SNPs for the 113 honey bee samples.
(XLSX)

S2 Table. Information content values of the initial 1183 SNP dataset estimated by the five selection methods (Weir & Cockerham's F_{ST}, Delta, informativeness (I_n), PCA and the F_{ST}-based outlier test) and for the four training datasets (I to IV). The SNPs are ordered from high to low information content. The top 48, 96, 144, 192 and 384 SNPs were included in the five reduced panels. SNPs marked with an asterisk (*) were excluded from the reduced panels because they were within a genetic distance <1 cM of other informative SNPs.
(DOCX)

S3 Table. Admixture proportion estimates inferred from the five AIMs panels (48-, 96-, 144-, 192-, 384-AIMs) and the initial 1183 SNP dataset for the holdout set. The holdout set consisted of 34 pure (training set) and 43 reserved individuals of *A. m. mellifera* and all reference individuals of *A. m. ligustica* (17) and *A. m. carnica* (19). * Samples marked with an asterisk (*) are of *A. m. mellifera* from protected populations (pure breeding for conservation purposes; see Pinto et al. 2014 [9] for details).
(DOCX)

S4 Table. *P*-values of Mann-Whitney pairwise several-sample-test. Values obtained from comparing individual admixture proportions estimated with the five AIMs panels (48-, 96-, 144-, 192-, 384-AIMs) and the initial 1183 SNP dataset using the holdout set.
(DOCX)

S5 Table. *P*-values of Mann-Whitney pairwise several-sample-test. Values obtained from comparing admixture proportions inferred from the five AIMs panels and the 1183 initial SNP dataset using the simulated set. The simulated set was generated with the program ONCOR (Kalinowski et al. 2007) using the function "simulate a single mixture". Ten populations, each with 100 genotypes, were simulated using different levels of C-lineage introgression (0, 1, 5, 10, 20, 30, 40, 50, 75, and 90%).
(DOCX)

Acknowledgments

We are deeply grateful to Andrew Abrahams, Bjørn Dahle, Gabriele Soland-Reckeweg, Gilles Fert, Lionel Garnery, Norman Carreck, Pilar de la Rúa, Raffaele Dall'Olio, and Romee Van der Zee for providing honey bee samples. DNA extractions and SNP genotyping were performed by Colette Abbey, with support from the TAMU Institute of Genomic Science and Society. An earlier version of the manuscript was improved by the constructive comments made by two anonymous reviewers.

Author Contributions

Conceived and designed the experiments: MAP IM DH. Analyzed the data: IM DH JC-G. Contributed reagents/materials/analysis tools: PK. Wrote the paper: MAP IM JSJ.

References

1. Dowling TE, Secor CL The role of hybridization and introgression in the diversification of animals. Annu Rev Ecol Evol Syst. 1997; 28: 593–619.
2. Nolte AW, Tautz D. Understanding the onset of hybrid speciation. Trends Genet. 2009; 26: 54–58.
3. Rhymer JM, Simberloff D. Extinction by hybridization and introgression. Annu Rev Ecol Evol Syst. 1996; 27: 83–109.

4. Allendorf FW, Luikart G. Conservation and the Genetics of Populations. 1st ed. Malden, Massachusetts: Blackwell Publishing; 2007.
5. Crane E. The World History of Beekeeping and Honey Hunting. 1st ed. New York: Routledge; 1999.
6. vanEngelsdorp D, Meixner MD. A historical review of managed honey bee populations in Europe and the United States and the factors that may affect them. J Invertebr Pathol. 2010; 103: 80–95.
7. Potts SG, Biesmeijer JC, Kremen C, Neumann P, Schweiger O, Kunin WE. Global pollinator declines: trends, impacts and drivers. Trends Ecol Evol. 2010; 25: 345–353. doi: 10.1016/j.tree.2010.01.007 PMID: 20188434
8. De la Rúa P, Jaffé R, Dall'Olio R, Muñoz I, Serrano J. Biodiversity, conservation and current threats to European honeybees. Apidologie. 2009; 40: 263–284.
9. Pinto MA, Henriques D, Chávez-Galarza J, Kryger P, Garnery L, van der Zee R, et al. Genetic integrity of the Dark European honey bee (*Apis mellifera mellifera*) from protected populations: a genome-wide assessment using SNPs and mtDNA sequence data. J Apic Res. 2014; 53: 269–278.
10. Meixner MD, Costa C, Kryger P, Hatjina F, Bouga M, Ivanova E, et al. Conserving diversity and vitality for honey bee breeding. J Apic Res. 2010; 49: 85–92.
11. Büchler R, Costa C, Hatjina F, Andonov S, Meixner MD, Le Conte Y, et al. The influence of genetic origin and its interaction with environmental effects on the survival of *Apis mellifera* L. colonies in Europe. J. Apic Res. 2014; 53: 205–214.
12. Ruttner F. Biogeography and Taxonomy of Honeybees. 1st ed. Berlin, Germany: Springer Verlag; 1988.
13. Hepburn HR, Radloff SE. (1998) Honey bees of Africa. Berlin, Germany: Springer. 370 p.
14. Engel MS. The taxonomy of recent and fossil honey bees (Hymenoptera: Apidae; *Apis*). J Hymenopt Res. 1999; 8: 165–196.
15. Sheppard WS, Meixner MD. *Apis mellifera pomonella*, a new honey bee subspecies from Central Asia. Apidologie. 2003; 34: 367–375.
16. Meixner MD, Leta MA, Koeniger N, Fuchs S. The honey bees of Ethiopia represent a new subspecies of *Apis mellifera—Apis mellifera simensis* n. ssp. Apidologie. 2011; 42: 425–437.
17. Garnery L, Cornuet JM, Solignac M. Evolutionary history of the honey bee *Apis mellifera* inferred from mitochondrial DNA analysis. Mol Ecol. 1992; 1: 145–154. PMID: 1364272
18. Garnery L, Solignac M, Celebrano G, Cornuet JM. A simple test using restricted PCR amplified mitochondrial DNA to study the genetic structure of *Apis mellifera* L. Experientia. 1993; 49: 1016–1021.
19. Arias MC, Sheppard WS. Molecular phylogenetics of honey bee subspecies (*Apis mellifera* L.) inferred from mitochondrial DNA sequence. Mol Phylogenet Evol. 1996; 5: 557–566. PMID: 8744768
20. Whitfield CW, Behura SK, Berlocher SH, Clark AG, Johnston JS, Sheppard WS, et al. Thrice out of Africa: ancient and recent expansions of the honey bee, *Apis mellifera*. Science. 2006; 314: 642–645. PMID: 17068261
21. Wallberg A, Han F, Wellhagen G, Dahle B, Kawata M, Haddad N, et al. A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee *Apis mellifera*. Nat Genet. 2014; 46: 1081–1088. doi: 10.1038/ng.3077 PMID: 25151355
22. Garnery L, Franck P, Baudry E, Vautrin D, Cornuet JM, Solignac M. Genetic diversity of the West European honey bee (*Apis mellifera mellifera* and *A. m. iberica*). I. Mitochondrial DNA. Genet Sel Evol. 1998a; 30: 31–47.
23. Garnery L, Franck P, Baudry E, Vautrin D, Cornuet JM, Solignac M. Genetic diversity of the West European honey bee (*Apis mellifera mellifera* and *A. m. iberica*). II. Microsatellite loci. Genet Sel Evol. 1998b; 30: 49–74.
24. Miguel I, Iriondo M, Garnery L, Sheppard WS, Estonba A. Gene flow within the M evolutionary lineage of *Apis mellifera*: role of the Pyrenees, isolation by distance and post-glacial re-colonization routes in the Western Europe. Apidologie. 2007; 38: 141–155.
25. Strange JP, Garnery L, Sheppard WS. Morphological and molecular characterization of the Landes honey bee (*Apis mellifera* L.) ecotype for genetic conservation. J Insect Conserv. 2008; 12: 527–537.
26. Oleksa A, Chybicki I, Tofilski A, Burczyk J. Nuclear and mitochondrial patterns of introgression into native dark bees (*Apis mellifera mellifera*) in Poland. J Apic Res. 2011; 50: 116–129.
27. Pinto MA, Muñoz I, Chávez-Galarza J, De la Rúa P. The Atlantic side of the Iberian Peninsula: a hotspot of novel African honey bee maternal diversity. Apidologie. 2012; 43: 663–673.
28. Uzunov A, Meixner MD, Kiprijanovska H, Andonov S, Gregorc A, Ivanova E, et al. Genetic structure of Apis mellifera macedonica in the Balkan Peninsula based on microsatellite DNA polymorphism. J Apic Res. 2014; 53: 285–285.

29. Muñoz I, Dall'Olio R, Lodesani M, De la Rúa P. Population genetic structure of coastal Croatian honeybees (*Apis mellifera carnica*). Apidologie. 2009; 40: 617–626.

30. Nedić N, Francis RM, Stanisavljević L, Pihler I, Kezić N, Bendixen C, et al. Detecting population admixture in the honey bees of Serbia. J Apic Res. 2014; 53: 303–313.

31. Jensen AB, Palmer KA, Boomsma JJ, Pedersen BV. Varying degrees of *Apis mellifera ligustica* introgression in protected populations of the black honeybee, *Apis mellifera mellifera*, in northwest Europe. Mol Ecol. 2005; 14: 93–106. PMID: 15643954

32. Soland-Reckeweg G, Heckel G, Neumann P, Fluri P, Excoffier L. Gene flow in admixed populations and implications for the conservation of the Western honey bee, *Apis mellifera*. J Insect Conserv. 2009; 13: 317–328.

33. Dreher K. Gedanken zum Neuaufbau des Zuchtwesens. Die Hessische Biene. 1946; 81: 62–64.

34. Maul V, Hähnle A. Morphometric studies with pure bred stock of *Apis mellifera carnica* from Hessen. Apidologie. 1994; 25: 119–132.

35. Jensen AB, Pedersen BV. Honey bee Conservation: a case story from Læsø island, Denmark. In: Lodesani M, Costa C, editors. Beekeeping and conserving biodiversity of honey bee. Sustainable bee breeding. Theoretical and practical guide. Hebden Bridge: Northern Bee Books; 2005. pp. 142–164.

36. Rortais A, Arnold G, Alburaki M, Legout H, Garnery L. Review of the DraI COI—COII test for the conservation of the black honeybee (*Apis mellifera mellifera*). Conserv Genet Res. 2011; 3: 383–391.

37. Weinstock GM, Robinson GE, Gibbs RA, Worley KC, Evans JD, Maleszka R, et al. Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature. 2006; 443: 931–949. PMID: 17073008

38. Chávez-Galarza J, Henriques D, Johnston JS, Azevedo JC, Patton JC, Muñoz I, et al. Signatures of selection in the Iberian honey bee (*Apis mellifera iberiensis*) revealed by a genome scan analysis of single nucleotide polymorphisms. Mol Ecol. 2013; 22: 5890–5907. doi: 10.1111/mec.12537 PMID: 24118235

39. Harpur BA, Kent CF, Molodtsova D, Lebon JM, Alqarni AS, Owayssc AA, et al. Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. P Natl Acad Sci USA. 2014; 111: 2614–2619. doi: 10.1073/pnas.1315506111 PMID: 24488971

40. Harpur BA, Minaei S, Kent CF, Zayed A. Admixture increases diversity in managed honey bees: reply to De la Rúa et al., 2013. Mol Ecol. 2013; 22: 3211–3215. doi: 10.1111/mec.12332 PMID: 24433573

41. Rosenberg NA, Li LM, Ward R, Pritchard JK. Informativeness of genetic markers for inference of ancestry. Am J Hum Genet. 2003; 73: 1402–1422. PMID: 14631557

42. Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, et al. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. Hum Mutat. 2009; 30: 69–78. doi: 10.1002/humu.20822 PMID: 18683858

43. Galanter JM, Fernandez-Lopez JC, Gignoux CR, Barnholtz-Sloan J, Fernandez-Rozadilla C, Via M, et al. Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. PLoS Genet. 2012; 8: e1002554. doi: 10.1371/journal.pgen.1002554 PMID: 22412386

44. Frantz AC, Pourtois JT, Heuertz M, Schley L, Flamand MC, Krier A, et al. Genetic structure and assignment tests demonstrate illegal translocation of red deer (*Cervus elaphus*) into a continuous population. Mol Ecol. 2006; 15: 3191–3203. PMID: 16968264

45. Wilkinson S, Archibald AL, Haley CS, Megens H-J, Crooijmans RPMA, Groenen MAM, et al. Development of a genetic tool for product regulation in the diverse British pig breed market. BMC Genomics. 2012; 13:580. doi: 10.1186/1471-2164-13-580 PMID: 23150935

46. Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, et al. Estimating African American admixture proportions by use of population-specific alleles. Am J Hum Genet. 1998; 63: 1839–1851. PMID: 9837836

47. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003; 164: 1567–1587. PMID: 12930761

48. Hoggart CJ, Shriver MD, Kittles RA, Clayton DG, McKeigue PM. Design and analysis of admixture mapping studies. Am J Hum Genet. 2004; 74: 965–978. PMID: 15088268

49. Pardo-Seco J, Martinón-Torres F, Salas A. Evaluating the accuracy of AIM panels at quantifying genome ancestry. BMC Genomics. 2014; 15: 543. doi: 10.1186/1471-2164-15-543 PMID: 24981136

50. Francis RM, Kryger P, Meixner M, Bouga M, Ivanova E, Andonov S, et al. The genetic origin of honey bee colonies used in the COLOSS Genotype-Environment Interactions Experiment: a comparison of methods. J Apic Res. 2014; 53: 188–204.

51. Bertrand B, Alburaki M, Legout H, Moulin S, Mougel F, Garnery L. MtDNA COI-COII marker and drone congregation area: An efficient method to establish and monitor honeybee (*Apis mellifera* L.) conservation centres. Mol Ecol Res. 2014.

52. Sambrook J, Fritsch EF, Maniatis T. Molecular Cloning: A Laboratory Manual. 2nd ed. Cold Spring Harbor: Cold Spring Harbor Laboratory Press; 1989.

53. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet. 2007; 81: 559–575. PMID: 17701901

54. Weir RJ, Cockerham CC. Estimating F-Statistics for the Analysis of Population Structure. Evolution. 1984; 38: 1358–1370.

55. Raymond M, Rousset F. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. J Heredity. 1995; 86: 248–249.

56. Foll M, Gaggiotti O. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics. 2008; 180: 977–993. doi: 10.1534/genetics.108.092221 PMID: 18780740

57. Hammer Ø, Harper DAT, Ryan PD. PAST: paleontological statistics software package for education and data analysis. Palaeontol Electron. 2001; 4(1): art. 4.

58. Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW, et al. PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet. 2007; 3: 1672–1686. PMID: 17892327

59. Anderson EC. Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. Mol Ecol Res. 2010; 10: 701–710. doi: 10.1111/j.1755-0998.2010.02846.x PMID: 21565075

60. Storer CG, Pascal CE, Roberts SB, Templin WD, Seeb LW, Seeb JE. Rank and order: Evaluating the performance of SNPs for individual assignment in a non-model organism. PLoS ONE. 2012; 7: e49018. doi: 10.1371/journal.pone.0049018 PMID: 23185290

61. Ozerov M, Vasemägi A, Wennevik V, Diaz-Fernandez R, Kent M, Gilber J, et al. Finding markers that make a difference: DNA pooling and SNP-arrays identify population informative markers for genetic stock identification. PLoS ONE. 2013; 8: e82434. doi: 10.1371/journal.pone.0082434 PMID: 24358184

62. Kalinowski ST, Manlove KR, Taper ML. ONCOR A computer program for Genetic Stock Identification; 2007. Available: Department of Ecology, Montana State University, Bozeman MT 59717. Accessed: http://www.montana.edu/kalinowski.

63. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009; 19: 1655–1664. doi: 10.1101/gr.094052.109 PMID: 19648217

64. Wright S. Evolution and the genetics of population, variability within and among natural populations. Chicago: University Chicago Press. 1978.

65. Wilkinson S, Wiener P, Archibald AL, Law A, Schnabel RD, McKay SD, et al. Evaluation of approaches for identifying population informative markers from high density SNP chips. BMC Genetics. 2011; 12: 45. doi: 10.1186/1471-2156-12-45 PMID: 21569514

66. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. 2005; 15(11): 1496–1502. PMID: 16251459

67. Albrechtsen A, Nielsen FC, Nielsen R. Ascertainment biases in SNP chips affect measures of population divergence. Mol Biol Evol. 2010; 27(11): 2534–254768. doi: 10.1093/molbev/msq148 PMID: 20558595

68. Meixner MD, Pinto MA, Bouga M, Kryger P, Ivanova E, Fuchs S. Standard methods for characterizing subspecies and ecotypes of *Apis mellifera*. J Apic Res. 2013; 52(4): 1–27.